DOCUMENT RESUME

ED 062 674 24 CG 007 246

AUTHOR Greeno, James G.
TITLE Progress Toward a Theory of Cognitive Structure. Final Report.
INSTITUTION Michigan Univ., Ann Arbor.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
BUREAU NO BR-9-0447
PUB DATE Jan 72
GRANT OEG-0-9-320447-4194(010)
NOTE 112p.

EDRS PRICE MF-$0.65 HC-$6.48
DESCRIPTORS Cognitive Development; *Cognitive Processes; *Individual Differences; *Instruction; Instructional Design; *Learning Processes; *Problem Solving; Teaching Procedures; Thought Processes

ABSTRACT

The purpose of this project was to develop new information and analyses that would contribute to development of a systematic understanding of cognitive structure, including its acquisition and utilization during problem solving. Experimental and theoretical work was done on three specific problems. (1) Studies of individual differences and effects of instructional variables in learning certain probability concepts have been conducted, giving information about aptitude x treatment interaction and about the effects of instructional procedure on structural outcomes of learning. (2) Analyses of performance in a transportations problem indicated that the cognitive process of solving the problem is considerably simpler than the external structure of the problem, and gives considerable doubt to prospects for inferring cognitive process directly from overt responses in problem solving. (3) Observarion of subjects' acceptance and rejection of conclusions showed that a previously noticed tendency toward induction of class membership is very general, and also led to a new hypothesis about the psychological rule of inference that corresponds to logical implication. (Author)

ED 062 674

PA 24

OE NCERD
CG
PRJ - 9-0447

Final Report

Project No. 9-0447
Grant No. OEG-0-9-320447-4194 (010)

James G. Greeno
The University of Michigan
Ann Arbor, Michigan 48104

PROGRESS TOWARD A THEORY OF COGNITIVE STRUCTURE

January 1972

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

Basic Research Branch

CG 007 246

Final Report

Project No. 9-0447
Grant No. OE6-0-9-320447-4194 (010)

Progress Toward a Theory of Cognitive Structure

James G. Greeno
The University of Michigan
Ann Arbor, Michigan

January, 1972

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

Basic Research Branch

Author's Abstract

The purpose of this project was to develop new information and analyses that would contribute to development of a systematic understanding of cognitive structure, including its acquisition and utilization during problem solving. Experimental and theoretical work was done on three specific problems. (1) Studies of individual differences and effects of instructional variables in learning certain probability concepts have been conducted, giving information about aptitude x treatment interaction and about the effects of instructional procedure on structural outcomes of learning. (2) Analyses of performance in a transportation problem indicated that the cognitive process of solving the problem is considerably simpler than the external structure of the problem, and gives considerable doubt to prospects for inferring cognitive process directly from overt responses in problem solving. (3) Observation of subjects' acceptance and rejection of conclusions showed that a previously noticed tendency toward induction of class membership is very general, and also led to a new hypothesis about the psychological rule of inference that corresponds to logical implication.

Preface

Research reported in this document was carried out by several graduate students, whose names appear with their contributions. Two of these students, Dennis E. Egan and John C. Thomas, Jr. held predoctoral fellowships from the National Science Foundation.

Regarding publications, Thomas' thesis has appeared as a technical report.

Thomas, J. C., Jr. An analysis of behavior in the hobbits-orcs problem. University of Michigan, Human Performance Center Technical Report No. 31, August, 1971.

Two papers based partly on project work will appear as chapters in forthcoming books.

Greeno, J. G. Utilization of cognitive structure in problem solving and reasoning. In K. Wexler (Ed.), Cognitive structure and behavior. National Academy of Sciences (in press).

Greeno, J. G. On the acquisition of a simple cognitive structure. In E. Tulving and W. Donaldson (Eds.), Organization of Memory. Academic Press (in press).

The articles by Egan and Greeno, and Greeno and Mayer in PART I have been submitted for publication to the Journal of Educational Psychology. The article by Thomas in PART II has been submitted to Cognitive Psychology. The article by Stokes in PART II has been submitted to the Quarterly Journal of Experimental Psychology.

TABLE OF CONTENTS

INTRODUCTION..... 1

PART I. STUDIES OF INDIVIDUAL DIFFERENCE AND INSTRUCTIONAL PROCEDURE IN ACQUIRING COGNITIVE STRUCTURE..... 3

Chapter I.1. Acquiring Concepts by Discovery and Rule Learning..... 7

Chapter I.2. Does ATI Involve Structural or Merely Quantitative Interaction?..... 28

References..... 49

PART II. UTILIZATION OF COGNITIVE STRUCTURE IN PROBLEM SOLVING AND REASONING..... 1

Chapter II.1. An Analysis of Behavior in the Hobbits-Orcs Problem..... 3

Chapter II.2. Paleologic: Relationships between Human Thought and Truth Functional Logic and the Predicate Calculus..... 28

References..... 54

List of Tables

Table I.1-1 Test Statistics in Analyses of Variance for Experiment I..... 13

Table I.1-2 Examples of the Six Types of Questions Used in Experiment II..... 18

Table I.1-3 Comparison of Mean Number of Errors and Time to Solve Instructional Problem on First and Second Day..... 19

Table I.1-4 Tests Statistics in Analyses of Variance for Experiment II..... 21

Table I.1-5 Analysis of Posttest Scores for Weighted Abilities Grouping..... 23

Table I.2-1 Illustrative Posttest Problems..... 33

Table I.2-2 Scores and Numbers of Subjects in Groups of Ss Formed by Aptitude Tests..... 34

Table II.1-1 Proportion Correct and Mean Latency: Experiment I..... 8

Table II.1-2 Proportions of Error Moves for Difficult States..... 9

Table II.1-3 Proportions of Moves from State 111..... 9

Table II.1-4 Proportion of Correct Moves in Various States as a Function of Subcondition..... 11

Table II.1-5 Average Number of Moves to Solve..... 12

Table II.1-6 Proportion of Correct Moves in Subconditions Control-First Half and Experimental-First Half, First Trial Only..... 12

Table II.1-7 Changes in Response Probabilities Over Trials..... 13

Table II.1-8 Frequency of 3-state Sequences at State III..... 14

Table II.1-9 Frequence of Move Types in Feedback and No-Feedback Conditions..... 16

Table II.1-10 Estimated Mean and Variance of Thinking Times for Conditions of Experiment II and III..... 19

Table II.2-1 Percentage Responses Consistent with Predicate Paleologic, '0', and Opposite to Predicate Paleologic.. 37

List of Figures

Fig. I.1-1. Schematic representation of instructional sequence..... 9

Fig. I.1-2. Measures of learning as functions of ability grouping in Experiment I..... 11

Fig. I.1-3. Measures of learning as functions of ability grouping in Experiment II..... 20

Fig. I.1-4. Plots of method x context interaction and method x item type interaction..... 22

Fig. I.2-1. Aptitude x treatment interaction for four tests..... 35

Fig. I.2-2. Interaction between instructional sequence and format of posttest items for conditions differing in use of review questions during instruction..... 37

Fig. I.2-3. Interaction between instructional sequence and content of posttest items for conditions differing in use of review questions during instruction..... 38

Fig. I.2-4. Interaction between instructional sequence and format of posttest items for sets of posttest items differing in content..... 39

Fig. I.2-5. Interaction between instructional sequence and type of posttest items for sets of posttest items differing in content..... 41

Fig. I.2-6. Interaction of familiarity with probability concepts and format of posttest items for groups of Ss differing in instructional sequence and use of review questions during instruction..... 42

Fig. I.2-7. Interaction of familiarity with probability concepts and type of posttest items for different instructional sequences..... 43

Fig. I.2-8. Interaction between use of systematic strategy in permutations test and type of posttest items for different instructional sequences..... 45

Fig. II.1-1. Search graph for the Hobbits-Orcs problem..... 6

Fig. II-1-2. Cumulative relative frequency of solution time for subcondition Control-First Half of Experiments I and II combined..... 21

Fig. II-1-3. Cumulative relative frequencies of solution time for subconditions Experimental-Second Half and Control-Second Half for Experiments I and II combined..... 23

INTRODUCTION

The goal of the research carried out in this project has been to contribute to progress toward a satisfactory theory of cognitive structure. The educational importance of deepening our understanding of cognitive structure hardly needs documenting. Increasing concern to communicate structural concepts has characterized the development of both the "new math" and the "new English." When educational objectives focus on communication of structural concepts, educational practice depends on assumptions about cognitive structure, and evaluation of educational success requires techniques for assessing changes in cognitive structure. Until very recently, little or no psychological research has been directed toward the development of rigorous theory to describe the salient properties of cognitive structure and its modification. This project was motivated by a belief that vigorous effort to achieve systematic understanding of cognitive processes with complex structure is timely, both in being needed and in being feasible with techniques that have become available recently.

Our effort to make progress in this general objective consisted of experimental and theoretical work on three specific problems. First, we selected a mathematical concept -- binomial probability -- and have conducted two experiments in this project investigating effects of individual differences and instructional variables on acquisition of this concept. Our results have led us to recognize an important theoretical variable which we now call the external connectedness of a cognitive structure. We also have obtained information that is helpful in understanding what abilities are needed by a student in order to successfully engage in discovery learning, and we have information suggesting a close similarity between discovery learning and receptive learning that emphasizes meanings of concepts in the structure.

The problem mentioned above involves processes of acquisition. The other two problems studied involve processes by which cognitive structure is utilized in problem solving and reasoning. The second problem studied in this project involved relationships between cognitive change and overt performance during problem solving. Like other aspects of the project, this study involved development of new methods of experimentation and analysis. The problem used in the experiment was a version of a transportation problem chosen partly because a successful computer simulation is available for comparison with performance of human subjects. Observation of human problem solving revealed new information about this problem, showing a difficulty that was not anticipated in earlier analyses. Analysis of times taken in solving parts of the problem indicated that the structure of the problem in the subject's cognitive process was considerably

simpler than the external structure of the problem, with the solution apparently involving about four cognitive stages compared with eleven steps in the external process. This result, combined with observations of transfer between different parts of the problem, led to the general conclusion that the important structural changes occurring during problem solving are not related in any simple way to the sequence of external moves made by the subject. This conclusion is of considerable methodological importance, since it implies that inferences about problem-solving processes based on sequences of observable responses will often be seriously misleading.

The third problem studied in the project involved development of a model of the deductive processes used by subjects in ordinary reasoning. It is well known that the rules of formal logic -- especially the definition of implication -- do not correspond to the ways in which people ordinarily draw conclusions from assumptions. But a systematic description of the operations actually used by people in deductive reasoning has not been available. The study of this problem conducted in the present project was based on data collected from subjects who judged whether certain conclusions followed from premises given. We obtained information indicating that subjects' inferences are based on a set of systematic cognitive operations that fail to meet the criterion of formal consistency but are adaptive in the sense of dealing meaningfully with ordinary experience in productive ways.

The three specific problems studied in this project are reported in the two parts of this document. Each part begins with a brief resumé of the work done on the problem, including a general description of results and conclusions. The main body of each part consists of research reports in the form submitted as journal articles. From the resumé given at the beginning of each part, readers should be able to decide whether they wish to read the full reports of research in the main body, which include details of procedure, data, and statistical analysis. References to literature are included in each part of the project report.

PART I

ACQUISITION OF PROBABILITY CONCEPTS

This part of the project report contains reports of three experiments in which concepts of probability were taught to college-age subjects. The goal of this research was to obtain new information about the effects of individual differences and instructional treatments, with the intent of identifying variables whose effects need to be taken into account in theory.

The first two experiments, conducted by Egan and Greeno, compared learning of probability concepts by discovery and rule methods. Unlike some earlier investigators, we obtained large and reliable aptitude × treatment interaction (ATI), but primarily with aptitude tests dealing with knowledge and skills directly involved in the learning tasks. A test of subjects' familiarity with general concepts of probability gave reliable ATI in both studies. A test of subjects' skill in arithmetic computation gave reliable ATI in the first study, where subjects were required to carry out computation, but had no relation with performance in the second study, where computations were carried out for the subjects as part of the CAI system in which the experiment was conducted. A test of subjects' use of a systematic strategy in generating permutations gave reliable ATI when the test required subjects to keep previous responses in memory, although nc relation with performance was obtained when subjects kept all responses on paper in front of them. Our results regarding a measure of general ability (the Mathematics Scholastic Aptitude Test score) fit with the general picture obtained in the literature -- a reliable ATI was obtained in one experiment and not in another

In every case where reliable ATI was obtained, the form of the interaction was that performance depended on aptitude more strongly when learning was by our discovery method than in the rule method. This supports the conclusion that adequate preparation in prerequisite concepts and skills is necessary for successful achievement in discovery learning, and less so for adequate performance in rule learning.

Analysis of posttest performance indicated interesting and important differences in the learning outcomes achieved with the two kinds of instruction. We did not test long-term retention or transfer to new learning tasks, but we did use posttest items varying in their similarity to those used during learning. Reliable treatment × posttest interaction (TPI) was obtained involving both the context in which test items were presented (formal variables vs. word problems) and the type of problem used (familiar types, problems that required an algebraic transformation, and Luchins problems where a direct answer is available that may be hidden if the formula is applied thoughtlessly). In both cases, subjects who learned by the rule method were much more successful with posttest items that were just like those used in learning than they were with problems that involved a changed

context or type of problem. Subjects who learned by the discovery method showed a much smaller difference between performance on different kinds of items. Since rule-learning subjects were superior on all items a strong conclusion is not warranted by this result, but the result suggests that discovery-learning subjects are able to apply what they learn to a wider variety of situations than rule-learning subjects. This tentative conclusion would encourage an hypothesis that cognitive structure acquired in discovery learning is more thoroughly integrated with other knowledge the subject had previously than is cognitive structure acquired by rule learning.

On the basis of this and other results obtained in our laboratory, we have concluded that an important variable regarding cognitive structure is the degree of external connectedness -- the extent to which a cognitive structure is integrated with other aspects of a person's knowledge. And we suspect that structure acquired by our discovery method has stronger external connectedness than structure acquired by our rule method. On the other hand, rule-learning subjects' superiority on familiar problems and problems stated in formal variables suggests that the structure acquired with rule learning has stronger internal connectedness -- that is, stronger connections among the concepts of the structure itself.

The third experiment reported in this part of the project report was conducted by Greeno and Mayer. This study compared instructional treatments that were largely expository in character, but that differed in sequencing of ideas and in emphasis on aspects of the material. One treatment began with the binomial formula and proceeded to explain its components, emphasizing the formula's use in calculation. The other treatment began with the component concepts in the binomial formula and proceeded to develop their relationships in the overall structure, emphasizing the meanings of concepts. Aptitude tests like those used in Egan and Greeno's study were used, except that the test of skill in arithmetic computation was replaced by a test of arithmetic concepts such as associativity and distributivity, including application to fractions, exponents, and factorials to provide direct relevance to the materials taught in the experiment.

The results of this experiment formed a pattern very similar to that of Egan and Greeno's findings. ATI was obtained with tests of familiarity with probability concepts and of systematic strategy in generating permutations. (The latter test was used in the form requiring subjects to remember previous responses.) Performance with concept emphasis instruction was much more strongly related to aptitude than was performance with formula emphasis. With the MSAT and the test of arithmetic concepts, we obtained main effects of aptitude but no ATI. These results confirm the hypothesis that predictions of performance based on aptitude tests will differentiate between instructional treatments only when the tests involve skills and concepts that are specifically involved in the learning task. This experiment also

suggests that familiarity with general concepts of arithmetic, unlike general concepts of probability, is not a specific prerequisite for learning the concept of binomial probability, at least as it is taught in our procedures.

In the posttest used by Greeno and Mayer, context and type of problem were varied as in Egan and Greeno's study (with two additional item types) and one additional variable was manipulated in the post-test. The added variable was the content of posttest problems: some problems dealt with the binomial probability of r successes in n trials, some with the joint probability of a sequence of events, and some with the combinatorial number of sequences having a given number of successes. Reliable TPI were obtained with context and content of posttest items, but not with type. The form of the interactions was the same as that found by Egan and Greeno, except that they were apparently disordinal, permitting stronger conclusions. Subjects taught with the sequence emphasizing calculation did better on problems involving formal variables and the whole binomial concept than concept-emphasis subjects, while subjects taught with the sequence emphasizing meanings of concepts excelled on word problems and problems involving components of the total concept. We conclude that teaching that emphasizes concepts, like discovery learning, provides an outcome of learning that has strong external connectedness, while teaching that emphasizes algorithmic calculation, like rule learning, results in cognitive structure with stronger internal connectedness but weaker external connectedness. A further suggestion in the results of this study is that a structure with strong external connectedness may provide a better basis for solving problems that require using a part of the structure -- we might say that a structure with stronger external connectedness may be easier to take apart in situations where only a component is needed.

The close similarity between results obtained by Egan and Greeno and those obtained by Greeno and Mayer suggest a close similarity between the structural consequences of the variables manipulated in the two studies. These results encourage the hypothesis that discovery learning and expository learning that emphasizes meanings of concepts are functionally equivalent. Both kinds of instruction lead to stronger integration of new structure with previous knowledge -- the factor we have come to call external connectedness.

An additional variable manipulated by Greeno and Mayer was the use of review tests during instruction. Some subjects had to correctly answer test questions as they went along in order to be allowed to proceed, while other subjects merely went through the instructional sequence, reviewing material when they felt they needed to. The inclusion of review questions largely eliminated the structural differences between the outcomes of learning. This is an important practical finding, since it suggests that the use of review questions mainly strengthens those aspects of an instructional treatment that

are relatively weak or ineffective. Perhaps without review questions subjects focus their attention on aspects of the material they perceive to be central, giving less attention to apparently peripheral topics, while review questions force adequate attention to be given to all aspects of the material, including those given relatively less emphasis in the instructional sequence.

The main positive results reported in this part have to do with two questions. The first is, "Which subjects learn more with which kind of instruction?" The answer is based on findings involving ATI. We find that when aptitude is measured with tests of skills and concepts directly involved in the learning task, subjects of higher aptitude learn more in discovery learning or in instruction that emphasizes meanings of concepts. The second question is, "What kind of learning outcome results from what kind of instruction?" The answer is based on TPI. We find that rule learning that emphasizes algorithmic calculation leads to a cognitive structure with strong internal connectedness that is apparently superior for solving problems of the kind used during instruction. Discovery learning or expository learning that emphasizes meanings of concepts leads to a cognitive structure with strong external connectedness that apparently can be applied over a wider range of problem situations and may be easier to apply in situations where only a component of the structure is needed.

We have had the opportunity to examine two additional questions in the results of these experiments. One is, "What kind of learning outcome is achieved by which subjects?" A positive finding regarding this question would involve aptitude × posttest interaction (API), particularly of a disordinal kind, that would indicate different structural outcomes achieved by subjects of different ability in the same instructional treatment. The other question is, "What kind of learning outcome is achieved by which subjects in which instructional treatment?" Here a positive finding would involve aptitude × treatment × posttest interaction (ATPI) of a kind indicating that structural differences produced by different instructional treatments were different at different levels of aptitude. We have not obtained API of ATPI of a kind that would support conclusions of structural difference related to student aptitude. Our best conclusion based on present evidence is that the instructional procedure determines what can be learned in the sense of cognitive structure that can be acquired, and student aptitude determines how much will be learned in a given instructional procedure. But student aptitude has little or nothing to do with what kind of structural outcome will be achieved in the learning process.

Chapter I.1

Acquiring Cognitive Structure by Discovery and Rule Learning

Dennis E. Egan and James G. Greeno

The University of Michigan

Understanding the effects of aptitude, instructional method, and their interaction (the aptitude-treatment interaction or ATI) is important in the study of learning and problem solving for at least three reasons. First, a thorough understanding of these effects may make it possible to assign Ss of differing ability to optimal instructional methods (Cronbach, 1967). Second, the process of acquiring cognitive structure can be analyzed in terms of the skills that are more or less relevant to success under different instructional methods. In this case, aptitude becomes a theoretical process variable (Melton, 1967). Third, the characteristics of cognitive structure acquired by different instructional groups can be inferred from group differences in terminal performance (Mayer & Greeno, in press).

Two experiments were performed to investigate the effects of aptitude and instructional method on learning concepts of probability.

Experiment I

Learning by discovery and learning by rule are contrasting instructional methods that appear important for applications and promising for analysis of process and structural distinctions. These methods have, in one form or another, been the focus of much research (Ausubel, 1961; Bruner, 1961; Corman, 1957; Gagné & Brown, 1961; Guthrie, 1968; Kittel, 1957; Shulman, 1970; Tallmadge, 1968; Wittrock, 1963). While studies have come to contradictory conclusions about the superiority of a discovery-type or a rule-type instructional method, there appears to be a consensus on the fundamental difference between learning by discovery and learning by rule. Subjects learning by discovery proceed by solving problems and generalizing with very little initial information. The task of the rule learner is to interpret initial information and apply it to problems. Other differences between the methods are probably not as essential.

A simple hypothesis suggests that skills involved in solving problems and generalizing are more important to success in learning by discovery than in learning by rule. This idea leads to the expectation of an ATI such that the skills of Ss learning by discovery should be strongly related to their performance while the skills of Ss learning by rule should be less strongly related to performance.

Available evidence appears to discredit this hypothesis. Tallmadge (1968) and Corman (1957) found no reliable ATI for groups of varying

ability learning by a discovery-type or a rule-type method. These studies used scores on tests of general ability as measures of aptitude. Recently Bracht (1970) surveyed ATI literature and reported that a disordinal ATI is more likely to be found if the tests of ability are specific to the learning task. Thus, the lack of evidence may be due to the use of tests of general ability. Moreover, an ATI found with a general aptitude would yield very little information about the processes of learning. The first experiment was performed in an attempt to achieve reliable ATIs in the expected direction, as well as to analyze the processes involved in learning by discovery and learning by rule.

Method

Materials -- Subjects were taught how to solve problems involving binomial probability by one of two different programmed texts. The texts were constructed by parsing an instructional binomial problem into a hierarchy of components. This instructional problem required finding the probability of three successes in five trials of rolling a die. Subjects advanced through the text by solving multiple choice problems concerning each component of the problem. The sequence is presented schematically in Fig. 1 where components are represented by their symbols in the formula. A correct answer allowed S to bypass lower level instruction on that particular component (Campbell, 1963), while an incorrect answer sent S into a remedial loop. Once the entire instructional problem was solved, S had to successively solve three criterion problems that changed the values of the instructional problem.

Subjects learning by rule were given the binomial formula and relevant definitions on the first page of the text. Thereafter, all questions and instruction were phrased in terms of the formula. Subjects learning by discovery were asked the same questions at each stage of the hierarchy as Ss learning by rule. However, the questions for the discovery group were phrased in ordinary English, as nontechnically as possible. For example, Ss learning by rule were asked to find the value of $p^r * q^{n-r}$ at the same point in the instructional sequence that Ss learning by discovery were asked to find the probability of a particular sequence of rolls. Definitions and notation for the variables were introduced to discovery Ss only after they had solved various parts of the instructional problem. Using the notation, Ss generalized their solutions to obtain parts of the formula. Discovery Ss never saw the entire binomial formula at once. Sequencing in the discovery and rule texts was identical.

Ability tests -- Tests of three abilities specific to binomial probability were administered. A test of probabilistic concepts consisted of 14 multiple choice questions concerning identification of the probabilities of single events, joint events, the nonoccurrence of events, the occurrence of either of two events, and the occurrence of

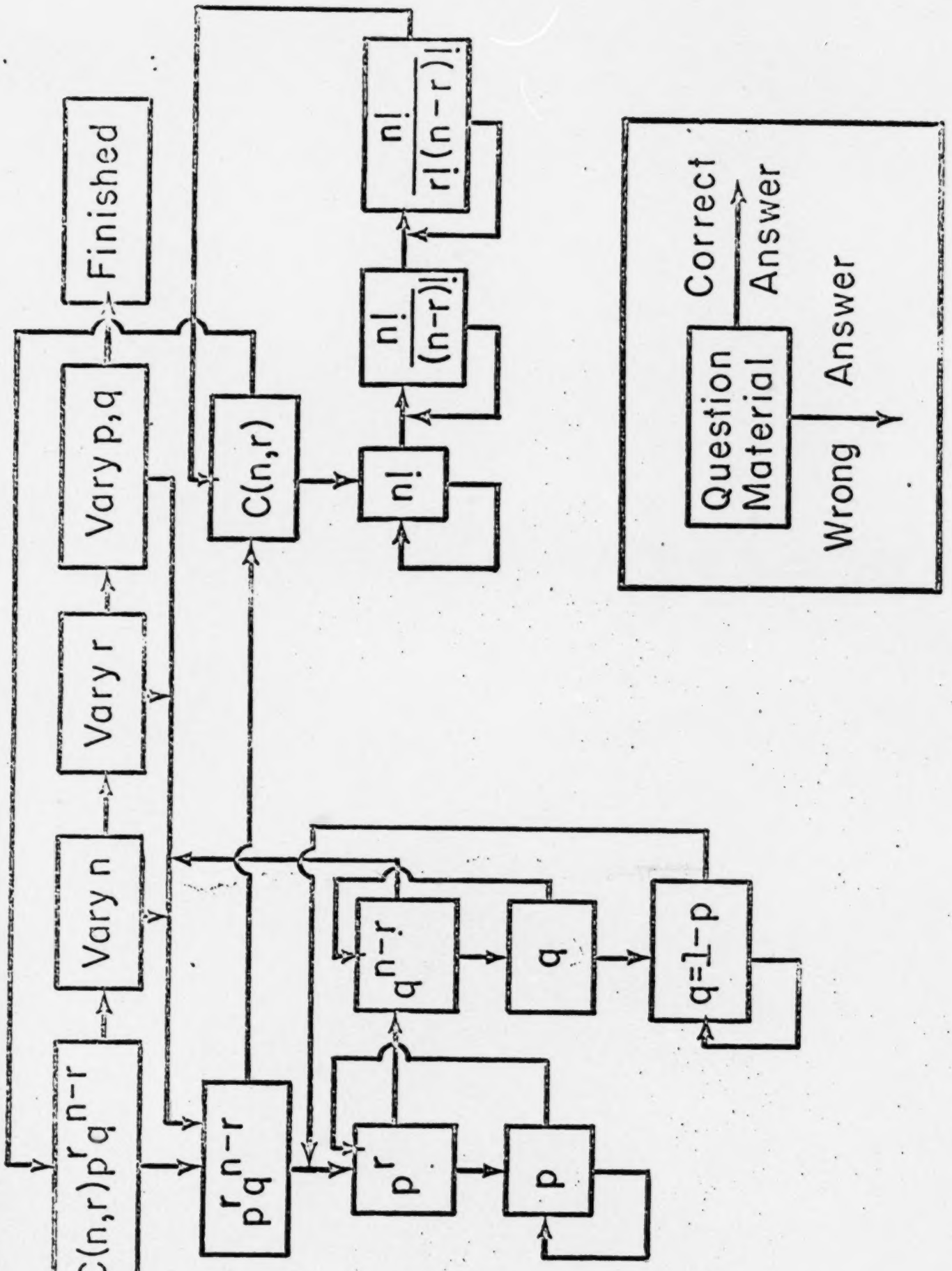


Fig. I.1-1. Schematic representation of instructional sequence.

simple sequences of events. A second test measured skill in the arithmetic operations necessary for calculating binomial probabilities. Eight problems were given involving computation of factorials, addition of fractions, and exponentiation of fractions. The third test was adapted from Leskow & Smock (1970). Subjects were asked to write out as many of the permutations of the digits 1234 as they could according to a plan that would exhaust all possibilities without repeating any. Scores were based on how closely S approximated one of two strategies: (1) holding initial digist constant and changing digits on the right, or (2) rotating the preceding permutation. The relevance of the first two tests to binomial probability is obvious. With regard to the permutations test, Piaget & Inhelder (1951) have hypothesized that a prerequisite for understanding probability is the ability to deal systematically with a set of possibilities. In discovering probabilistic concepts, the ability to count the elements of an outcome space seems especially important. To obtain measures of general ability, Ss were asked to report their scores on the Mathematical Scholastic Aptitude Test (MSAT).

Procedure -- Subjects were given the pretests and then the programmed texts were handed out at random. When S completed the programmed booklet he was given a 5-min break before beginning the posttest. The posttest consisted of ten binomial questions involving different situations.

Subjects -- A total of 57 Ss (male and female) from the University of Michigan paid subject pool participated in the experiment, 29 in the discovery group and 28 in the rule group. Up to five Ss served in each experimental session.

Measures of Learning -- For each S three measures of learning were obtained: the number of errors made in answering the multiple choice problems in the programmed text, the amount of time taken to complete the instructional sequence correctly, and the proportion of errors made on the posttest.

Results

Scores on the permutations test did not account for a significant portion of variance for any of the three measures of learning. This test was excluded from further analyses. For the remaining three abilities, Ss were divided into three groups approximately equal in size on the basis of each test score.

Of the 57 Ss 43 provided their MSAT scores. The range was 419 to 774. Low scoring ($\leq$ 599; $N_D = 5$, $N_R = 8$), Intermediate (600 to 699; $N_D = 8$, $N_R = 8$), and High scoring ($\geq$700; $N_D = 9$, $N_R = 5$) were formed. The first column of Fig. 2 shows the relationship between MSAT scores and the three measures of learning.

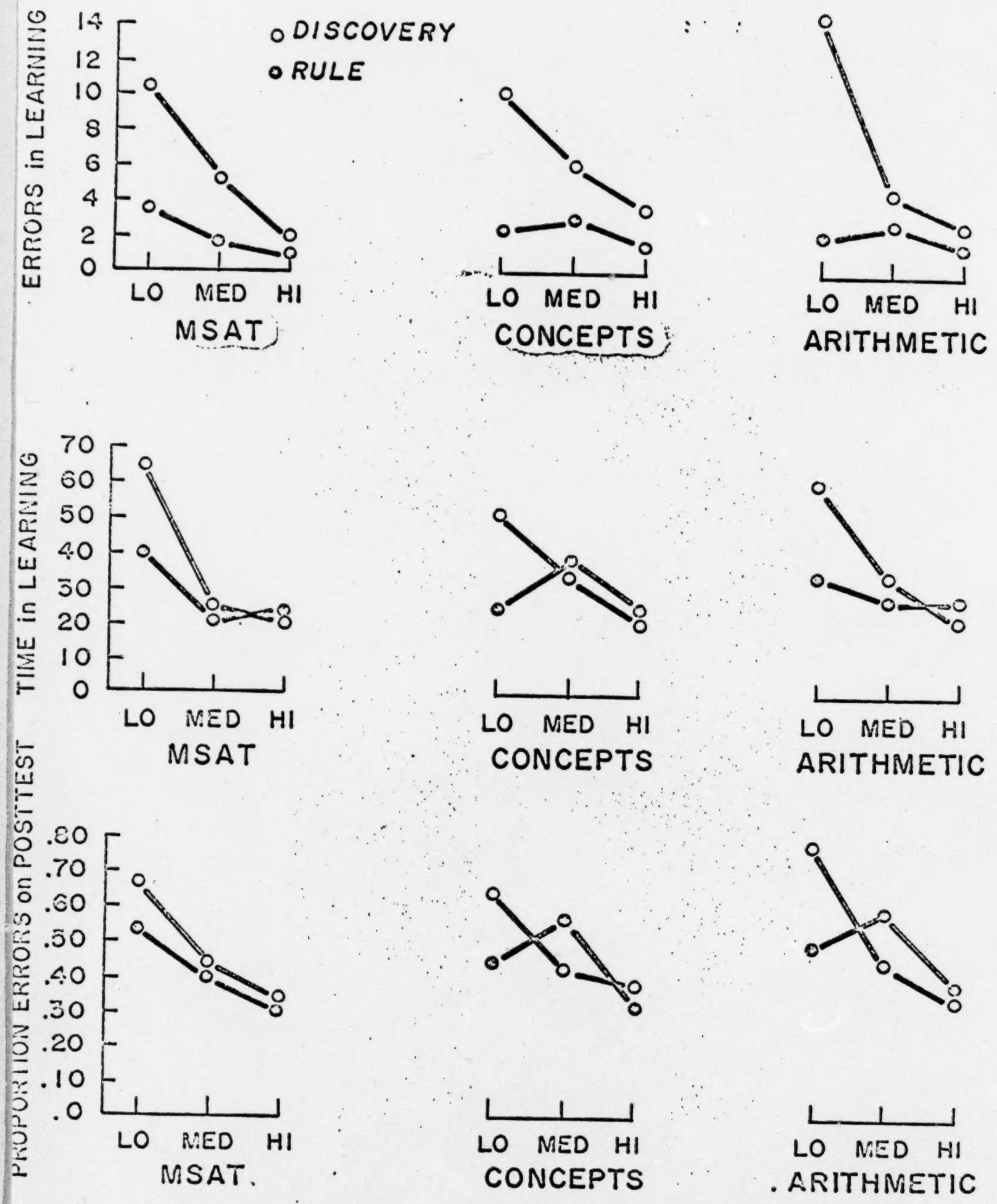


Fig. I.1-2. Measures of learning as functions of ability grouping in Experiment I.

Scores for the 57 Ss on the 14 item test of probabilistic concepts yielded a range of 5 to 14 correct. Low scoring ($\leq$ 10 correct; $N_D = 10$, $N_R = 6$), Intermediate (11 or 12 correct; $N_D = 8$, $N_R = 10$, and High scoring (13 or 14 correct; $N_D = 11$, $N_R = 12$) groups were formed. The middle column of Fig. 2 shows the results of the concepts grouping for all Ss.

Arithmetic operations scores ranged from 0 to 8. The sample was divided into Low scoring ($\leq$ 4 correct; $N_D = 6$, $N_R = 8$), Intermediate (5 to 7 correct; $N_D = 11$, $N_R = 7$), and High scoring (8 correct; $N_D = 12$, $N_R = 13$) groups. The third column of Fig. 2 shows the results when skill with arithmetic operations was used as the ability criterion.

Table 1 gives the results of analyses of variance for the various combinations of ability criteria and measures of learning.

Discussion

Several sets of findings are of psychological interest. First, consider overall differences due to instructional method. Subjects committed more errors in learning by discovery than in learning by rule. This difference is a straightforward result of the difference in methods, since the discovery method required Ss to first solve problems then infer principles from the problems. However, there was not a reliable difference between the two methods in time spent in learning. This finding suggests that there was not a substantial difference in the overall difficulty of the two teaching programs. The lack of a main effect due to method on the posttest suggests that there was no reliable difference in the effectiveness of instruction.

The differences among ability groups for all analyses were highly significant ($p < .01$). In every case, the groups scoring higher on the test of ability performed better on the measures of learning. Thus the tests of concepts and arithmetic operations as well as the MSAT measured characteristics relevant to the learning task.

The main point of the experiment was to test the hypothesis that skills involved in solving problems and generalizing are more important to success in learning by discovery than in learning by rule. Reliable ATIs were obtained in seven of the nine analyses, all in the expected direction. Thus the hypothesis was supported.

Specifically, from the graphs of errors in learning in Fig. 2, it is apparent that all three groups of Ss learning by rule made few errors, but groups of Ss learning by discovery were systematically ordered. The abler discovery Ss made fewest errors while the intermediate and low ability groups made progressively more errors. The same general pattern of results was obtained in analyses of time spent in learning.

Table I.1-1

Test Statistics in Analyses of Variance for Experiment I

Measure of Learning	Test of Ability	Ability Main Effect	Method Main Effect	Interaction Effect (ATI)
Errors in Learning	Arithmetic	$F(2,51)=9.46^{***}$	$F(1,51)=21.98^{***}$	$F(2,51)=11.99^{***}$
	Concepts	$F(2,51)=6.37^{***}$	$F(1,51)=14.77^{***}$	$F(2,51)=1.33$
	MSAT	$F(2,37)=8.27^{***}$	$F(1,37)=10.57^{***}$	$F(2,37)=6.65^{***}$
Time in Learning	Arithmetic	$F(2,51)=7.87^{***}$	$F(1,51)=1.59$	$F(2,51)=3.79^{**}$
	Concept	$F(2,51)=7.18^{***}$	$F(1,51)=1.59$	$F(2,51)=4.37^{***}$
	MSAT	$F(2,37)=19.99^{***}$	$F(2,37) < 1.00$	$F(2,37)=5.44^{***}$
Errors on Posttest	Arithmetic	$F(2,51)=6.97^{***}$	$F(1,51) < 1.00$	$F(2,51)=3.12^{***}$
	Concept	$F(2,51)=6.89^{***}$	$F(1,51) < 1.00$	$F(2,51)=3.72^{**}$
	MSAT	$F(2,37)=5.56^{***}$	$F(1,37) < 1.00$	$F(2,37) < 1.00$

$^{***}p < .01$

$^{**}p < .05$

$^{*}.10 > p > .05$

Finally, consider the ATI on the posttest. Consistent with Corman (1957) and Tallmadge (1968), there was no evidence of an interaction between instructional method and general ability as measured by the MSAT. However, interactions were found between the methods used and the tests that measured abilities specifically involved in the learning task. The effect was at least marginally significant for both the test of concepts and the arithmetic test.

Knowledge of probabilistic concepts,and arithmetic operations was more important to success in learning by this version of discovery than this version of rule. To that extent there is some clue as to the difference between the process of learning by discovery and the process of learning by rule. If acquisition of concepts by discovery involves more problem solving and generalizing activity than does learning by rule, it would be expected that the learning outcomes produced by the two methods might differ. Since the set of problems on the posttest was not generated in any systematic fashion, little can be said concerning the characteristics of the cognitive structure produced by each method of instruction.

A second experiment was performed to replicate the obtained ATIs and to extend understanding of what is acquired under each type of instruction by means of a systematic transfer analysis.

Experiment II

Katona (1940) found that meaningful learning allows Ss to solve problems in a variety of circumstances. If Ss discovered the principle of solving a set of problems, they performed better on tests of long-term retention and transfer than Ss who had memorized and practiced a rule for solving the problems. On the other hand, when tested immediately on problems very similar to the instructional materials, Ss who had learned by memorizing and drill performed better.

Other reported differences in retention and transfer between Ss learning by discovery or learning by rule have been inconsistent (e.g., Kittel, 1957; Guthrie, 1968; Wittrock, 1963). The diversity of results is probably due in part to the diversity of instructional materials and instructional methods.

In one study that used instructional materials and methods similar to those in the present study, Gagné and Brown (1961) gave three groups of Ss programmed instruction in the summation of algebraic series. The groups of interest were the rule-example group and the guided discovery groups which roughly correspond to the rule and discovery groups in the present study. While all three instructional methods produced savings in time spent in relearning ((a measure of retention), the guided discovery group showed the highest proficiency in solving problems on a posttest (a measure of transfer).

Results of Experiment I indicated that there was no overall difference between the discovery and rule groups in number of problems solved on the posttest. Since a rather haphazard selection of problems was used, the discovery method might have produced better performance on some types of problems with the rule method producing better performance on other types of problems.

How might instructional method affect performance on various types of problems? The answer depends on the characteristics of the cognitive structure produced by each instructional method. One hypothesis is that the problem solving and generalizing activity required of Ss learning by discovery produces greater integration of new information into existing cognitive structure. Because Ss learning by discovery think about and solve problems before being given an algorithm, they understand the material in a more meaningful way (Katona, 1940) than Ss learning by rule. Subjects learning by discovery thus acquire new structural links between concepts already known, rather than first representing concepts by notation and then memorizing relations among coded variables.

If this hypothesis were true, then the difference in performance between fairly direct problems and problems requiring interpretation (in the sense of relating what was known previously to the principle recently learned) should be greater for Ss learning by rule than for Ss learning by discovery. Specifically, on posttest problems that are posed in terms of components of the formula, performance of Ss learning by rule should be relatively better than on word problems because word problems require more interpretation. Moreover, on problems on the posttest that can be solved by directly applying the rule, Ss learning by rule should perform relatively better than on problems that must first be transformed to apply the rule, or that cannot be solved by using the rule. If the structure acquired by Ss learning by discovery is well integrated then the performance of those Ss on a posttest should be less affected by changes in the amount of interpretation necessary.

Method

Materials -- Subjects were taught how to solve problems involving joint probability (e.g., finding the probability of a particular sequence of successes and failures) and by means of programmed instruction similar to the first half of the texts used in Experiment I. The instructional procedures differed from those in the first experiment in several important ways. First, a Computer Assisted Instruction (CAI) system was used instead of a programmed text. Subjects sat in booths equipped with keyboards and display screens and responded to questions by typing in answers. Second, Ss had to calculate and enter numerical answers rather than choose among a set of possible responses. Third, at all times Ss had several options available. Subjects could always

return to a frame that summarized the instructional problem; they could at any time get out of an instructional loop and attempt to solve the instructional problem; they could use a programmed arithmetic calculator for any difficult computations. Additionally, Ss learning by rule could return to a frame defining all the variables at any time. Finally, Ss learning by discovery were not exposed to the formula or definitions until the second day of the experiment.

Ability tests -- Tests were again given in conceptual, arithmetic, and permutation skills, but each test was modified somewhat from the first experiment. The test of probabilistic concepts consisted of eight questions concerning identification of the probability of single events, occurrence of either of two events, occurrence of joint events, and nonoccurrence of events. The test of arithmetic operations consisted of eight problems involving addition, subtraction, multiplication, and exponentiation of fractions. The permutation task was changed so that after S types in a permutation, his display screen was erased, leaving only the last acceptable permutation he wrote. This procedure is more similar to that used by Leskow & Smock (1970). Permutations were scored for the strategy of holding digits constant from the left. MSAT scores were again obtained as measures of general mathematical ability.

Procedure -- On the first day of the experiment, Ss were randomly assigned to the discovery or rule group. They then received instruction in the use of the CAI equipment, and were given the ability tests followed by the instructional problem which concerned finding the probability of a particular sequence of successes and failures in rolling a die. Subjects returned 24 hours later and again had to solve the instructional problem. Scores on solving the instructional problem were used to measure retention. Following the instructional problem, all Ss had to write out the formula for joint probability, $p^r * q^{n-r}$, once correctly. For Ss learning by discovery, this task required inferring the formula from their solution of the instructional problem. For Ss learning by rule, the task simply required giving the formula from memory as it had already been presented. Once Ss wrote out the formula correctly, they went on to the set of criterion problems. The posttest immediately followed the last criterion problem.

Transfer Design -- The posttest consisted of 18 problems, three of each of six types in a 2x3 design. The first factor was problem-context. Half the problems were word problems, half were posed in terms of the components of the formula. The second factor was problem-type and involved the amount of transformation necessary before the joint probability formula could be applied. Familiar problems were similar to the instructional and criterion problems in that all values necessary to solving the joint probability formula were explicitly stated and the formula could be directly applied to obtain a solution. Transformed problems did not state all values of the formula explicitly.

Instead, the S was required to obtain some of them by simple calculation. The third type of problem was called a Luchins problem (Luchins, 1942). These problems had very direct solutions, but were not solvable by direct application of the rule learned. An example of each of the six types of problems is given in Table 2. The problems were randomized at the start of each session.

Subjects -- A total of 72 Ss (male and female) from the University of Michigan paid subject pool participated in the experiment, 36 in each instructional group. The CAI system was set up to handle up to five Ss in a single session.

Measures of Learning -- For each S separate scores were obtained for errors made on questions in the programmed instruction and time in learning. For problems on the posttest, the overall proportion of errors made and the time spent in solving each problem were obtained for each S.

Results

Analysis of the relearning concerned comparing the errors and time to solve the instructional problem on the first and second day. Table 3 shows that Ss learning either by discovery or by rule solved the instructional problem on the second day in less time and with fewer errors than on the first day. Since so few Ss made any errors at all on the second presentation of the instructional problem, the partial errors and time scores were not analyzed for effects of ability. Instead, scores on the instructional problem for the first and second days were combined with errors and time taken to give the formula and solve the criterion problems. These summed scores of time and errors were used in all further analyses of learning.

Scores on the test of arithmetic operations were not strongly related to any of the measures of learning. The test was excluded from further analyses. On the basis of each of the remaining three abilities, Ss were divided into three groups of approximately equal size.

Of the 72 Ss, 65 provided their MSAT score. The range was 450 to 800. Low scoring ($\leq$ 599; N_D = 10, N_R = 10), Intermediate (600 to 699; N_D = 12, N_R = 15), and High scoring ($\geq$ 700; N_D = 11, N_R = 7) groups were formed. The first column of Fig. 3 shows the relationship between MSAT scores and three measures of learning (overall errors, overall time in learning, proportion of errors on the posttest).

Scores on the test of probabilistic concepts yielded a range of 0 to 8 correct. Subjects were grouped into Low (0 to 5; N_D = 11, N_R = 9), Intermediate (6 or 7; N_D = 17, N_R = 19), and High scoring groups (8 correct;

Table I.1-2

Examples of the Six Types of Questions Used in Experiment II

Word Questions

Familiar: A die has five spots on one of its six sides, and other numbers on the other sides. If you roll it ten times, what is the probability of getting three fives followed by seven other numbers?

Transformed: If you bet on 2 of 38 numbers in a game of roulette, you win only if one of those numbers is rolled. If you make such a bet, what is the probability of winning on the first two rolls and losing on the next three?

Luchins: You play a game five times in which the probability of winning each time is .17, and the probability of winning three games out of five is .32. What is the total number of successes plus the total number of failures?

Formula Questions

Familiar: R=2, N-R=4, P=1/5, Q=4/5. What is the joint probability?

Transformed: N=7, R=2, P=.31. What is the joint probability?

Luchins: Joint Probability = 15/128, N=5, P=.25, Q=.75. What is the value of R + (N-R)?

N_D = 8, N_R = 8). The middle column of Fig. 3 shows the results of grouping by scores on the test of concepts.

Table I.1-3

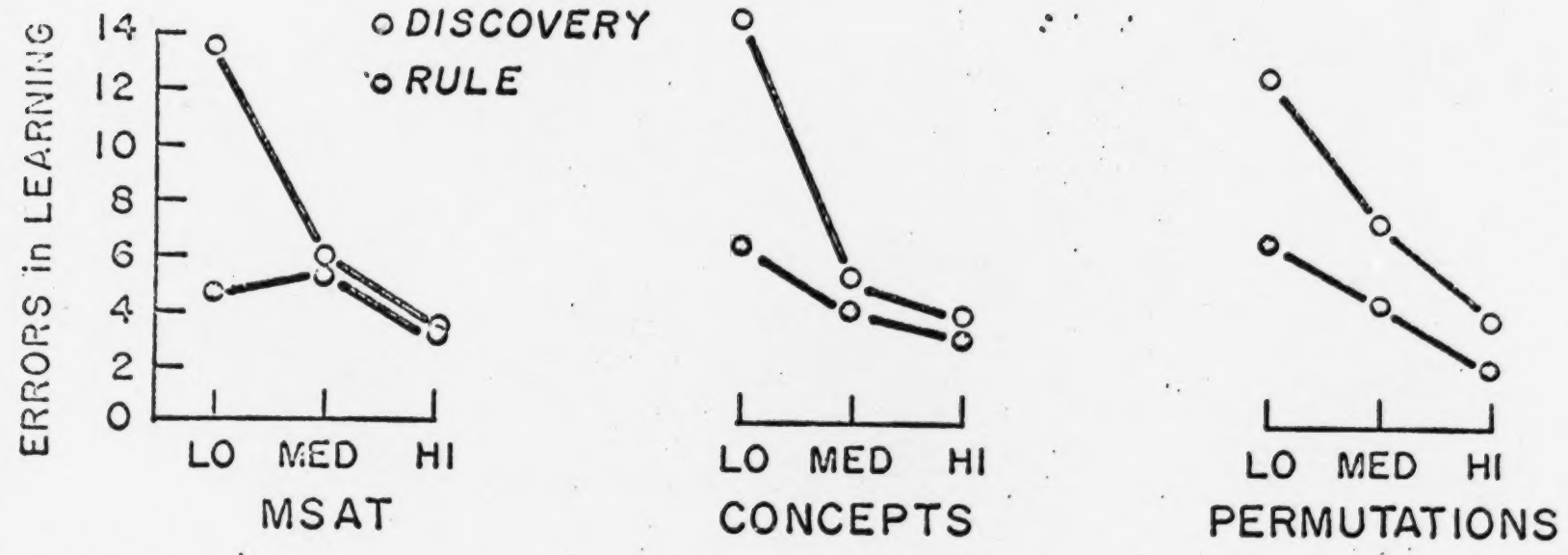
Comparison of Mean Number of Errors and Time to Solve Instructional Problem on First and Second Day

Group	Measure	First Day	Second Day	F
Discovery	Errors	3.6	0.6	14.92***
Rule	Errors	3.1	0.3	18.54***
Discovery	Time (min)	6.8	1.8	25.51***
Rule	Time (min)	11.6	2.7	61.28***

Scoring for the strategy of generating permutations by the number of digits held constant from the left gave a range of 1 to 32, the maximum score possible. Groups of Low ($\leq$ 11; N_D = 12, N_R = 14), Intermediate (11 to 29; N_D = 12, N_R = 10), and High (30 to 32; N_D = 12, N_R = 12) ability were formed. Results are presented in the last column of Fig. 3. Table 4 summarized the analyses of variance for all combinations of ability, instructional method and measure of learning.

Performance on the different kinds of posttest problems of Ss in the two conditions is graphed in Fig. 4. Data from the posttest were analyzed by means of a 2X3X2X3 analysis of variance for each ability grouping. Instructional method and aptitude level were between subject variables, and those results are incorporated in Fig. 3 and Table 4. Problem-context and problem-type were within-subject variables. As analyses of the posttest data for all three abilities followed the same general pattern, a weighting system was devised so that each score (concepts, permutations, MSAT) contributed about equally to the variance of a weighted abilities score.

Weighted Score = Concepts Score + $\frac{\text{Permutation Score}}{6}$ + $\frac{\text{MSAT}}{44}$. The full analysis based on the weighted abilities score is given in Table 5.



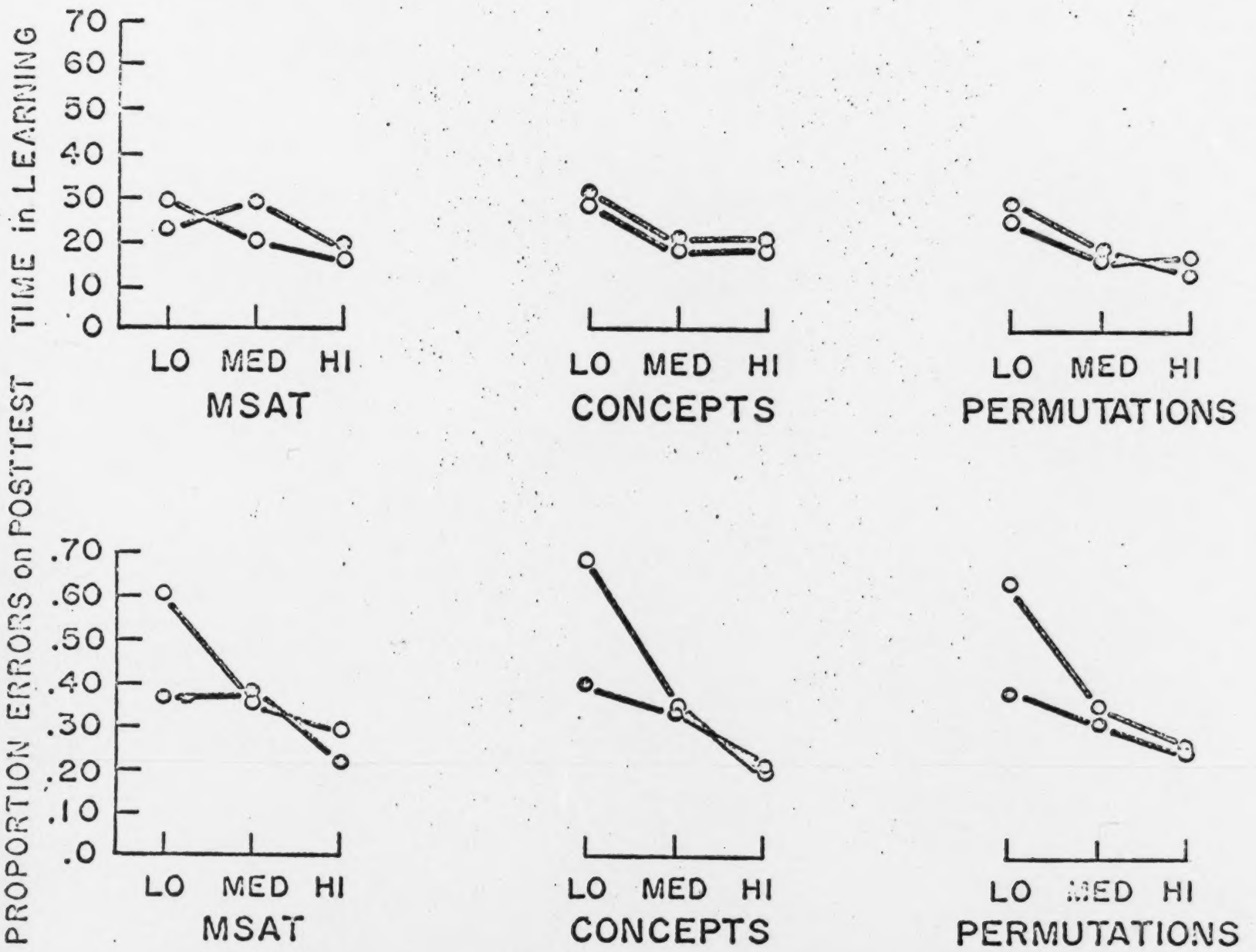


Fig. I.1-3. Measures of learning as functions of ability grouping in Experiment II.

Table I.1-4

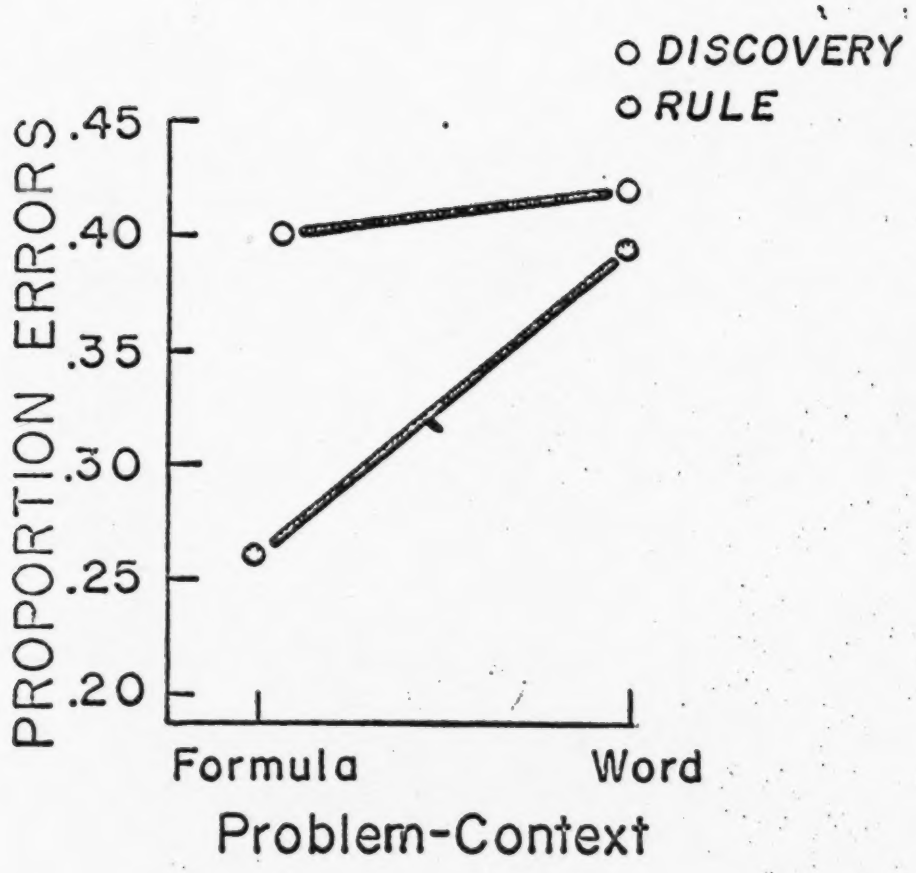
Tests Statistics in Analyses of Variance for Experiment II

Measure of Learning	Test of Ability	Ability Main Effect	Method Main Effect	Interaction Effect (ATI)
Errors in Learning	Permut.	F(2,66)=10.90***	F(1,66)=8.06***	F(2,66)=1.56
	Concept	F(2,66)=15.26***	F(1,66)=9.18***	F(2,66)=3.54**
	MSAT	F(2,59)=7.80***	F(1.59)=6.24**	F(2,59)=6.28***
Time in Learning	Permut.	F(2,66)=8.57***	F(1,66)=1.17	F(2,66) < 1.00
	Concept	F(2,66)=7.69***	F(1,66)=1.15	F(2,66) < 1.00
	MSAT	F(2,59)=5.02***	F(1,59)=1.22	F(2,59)=3.39**
Errors on Posttest	Permut.	F(2,66)=9.83***	F(1,66)=3.76*	F(2,66)=3.23**
	Concept	F(2,66)=17.76***	F(1,66)=4.47**	F(2,66)=4.04**
	MSAT	F(2,59)=8.90***	F(1,59)=1.44	F(2,59)=3.26**

***p < .01

**p < .05

*.10 > p > .05



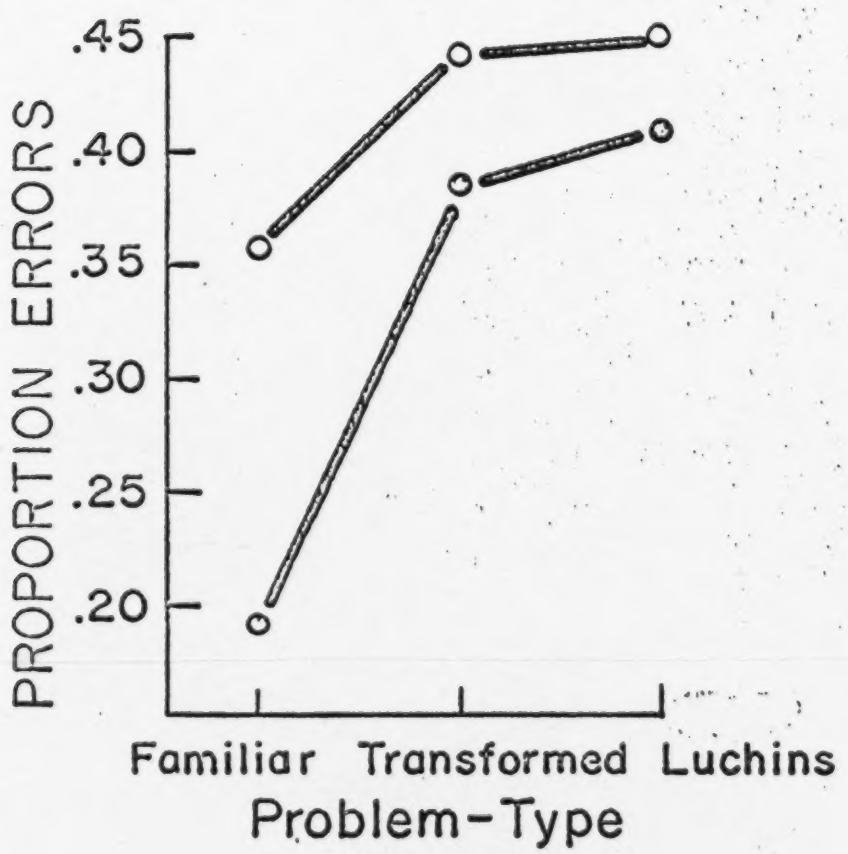


Fig. I.1-4. Plots of method x context interaction (top graph) and method x item type interaction (lower graph).

Table I.1-5

Analysis of Posttest Scores for Weighted Abilities Grouping

Source	SS	df	MS	F
A: Ability	89.43	2	44.72	31.94***
B: Method	7.78	1	7.78	5.56**
A x B	8.92	2	4.46	3.19**
Error (a)	92.30	66	1.40	
C: Problem-Context	5.79	1	5.79	9.98***
D: Problem-Type	19.09	2	9.54	16.45***
A x C	1.03	2	.52	.90
A x D	7.98	4	2.00	3.45**
B x C	3.34	1	3.34	5.76**
B x D	3.34	2	1.67	2.88*
C x D	7.03	2	3.52	6.07***
A x B x C	1.27	2	.64	1.10
A x B x D	.00	4	.00	.00
A x C x D	13.75	4	3.44	5.93***
B x C x D	2.57	2	1.28	2.21
A x B x C x D	3.51	4	.88	1.52
Error (b)	191.97	330	.58	
Total	459.10	431		

***$p < .01$

**$p < .05$

*$.10 > p > .05$

Discussion

One goal of studying aptitude and instructional variables is to be able to assign Ss of varying ability to optimal instructional methods. The present results suggest that Ss lacking in skills necessary to solve problems may learn more efficiently when instructed by techniques requiring interpretation and application of a rule. By every measure, Ss low in relevant abilities performed better when instructed by the rule method. That the rule method used in this study was not inherently better can be inferred from two results found in Experiment I and replicated in Experiment II. First, while Ss learning by discovery did generally make more errors on the teaching program, they still managed to learn the material in about the same amount of time as Ss learning by rule. Results in Table 3 indicate that Ss learning by discovery did not make more errors or take as much time as Ss learning by rule in solving the instructional problem. The extra time and errors were incurred when discovery Ss had to infer the formula and their solutions and apply it to the criterion problems. Second, there was little difference between instructional groups in overall performance on the posttest. The apparent method main effect in the analyses in Tables 4 and 5 was largely due to the simple effect of method for low-ability Ss.

A second goal of the present study was to describe the differences in the process of acquiring cognitive structure by discovery and rule. The fact that real differences exist was supported again in Experiment II where reliable ATIs were obtained in six of the nine tests, all in the expected direction. In Experiment I the discovery method required the availability of relevant probabilistic concepts and computational skills to a greater degree than the rule method. In Experiment II where Ss were given arithmetic calculators, computational skill was unrelated to performance, but the discovery method required conceptual ability and the ability to solve problems in a systematic way to a greater degree than the rule method.

Analysis of the differences in the process of acquiring cognitive structure might begin by identifying the component processes involved in learning under each method. First, consider the rule method. To solve parts of the instructional problem, a subject might carry out the following steps, not necessarily in a serial fashion.

1. Read the problem text.

2. Select information from the text pertaining to the values of relevant variables, and co-ordinate this information to the coded representations of varialbes in memory. For example, from the phrase, "the chances of success were 1/4," Ss could extract information in the form, "p = .25".

3. Select a rule or formula for using the variables whose values have been taken from the text. This might be looked up in available information, or retrieved from memory.

4. Perform any transformations needed to make the rule applicable to the information.

5. Calculate the answer.

Since the learning by rule did not greatly involve conceptual and other skills, individual differences in these skills were not associated with differences in performance. On the other hand, a measure of working memory, for example the ability to memorize, transform and apply formulas, might be related to success in learning by rule.

Now consider the discovery method. In the discovery method, Ss had to solve the instructional problem without first being given an algorithm. A discovery S might carry out the following steps:

1. Read the problem text.

2. Interpret the information in the problem in relation to concepts that are understood. The discovery method did not provide a well-specified list of variables as did the rule method. Therefore, interpretation of information in the discovery method probably had more of the properties of understanding a sentence than in the rule method, and less of the character of filling in values of variables in a list.

3. Search for or systematically generate relationships among concepts used in the problem, particularly relationships that seem to move in the direction of relating the given information with the unknown. This is the kind of process that has been investigated in classical studies of problem solving such as those of Duncker (1945), Polya (1965), and Wertheimer (1959). Subjects might find relationships that involved their understanding of the concepts in the problem, or they might apply a more general relational structure that fit the needs of the problem, or they might find a set of concepts in memory whose relationships seem to provide an analogy to the situation in the problem.

4. Carry out any calculations needed to obtain the answer. This process may well entail a great deal of computational ability, since no algorithm is present to relate specified variables and operations in a compact way.

Since the process of learning by discovery requires conceptual, systematizing and other skills, individual differences in these skills led to similar differences in performance.

Given these distinctions in the process of acquisition, it follows that there are differences in the learning outcomes of the two instructional groups. The results pertinent to this question involve the interactions of method and the two transfer variables appearing in Table 5 and graphed in Fig. 4. Both two-way interactions involving instructional method and transfer were at least marginally significant. While the overall performance of Ss learning by discovery is depressed because of the low ability group, Ss learning by rule showed a much greater decrement in performance on problems requiring more interpretation. The difference between percentage of formula and word problems solved was 13% for the rule group compared to 3% for the discovery group. Differences between Familiar and Luchins problems solved correctly were 22% for the rule group and 9% for the discovery group. These trends were present at all ability levels, although the curves for the two instructional methods crossed only in the high and intermediate ability groups. The average time taken to solve the six types of problems, given a correct solution, was also computed for each instructional group. These results are difficult to analyze because of missing data, but in general show the same method-transfer interactions.

These data indicate that the result of learning by discovery is a well integrated dognitive structure. Subjects can solve problems that require relating what they knew previously to the principle learned about as well as problems that require direct application of the principle. This feature of cognitive structure has been termed "external connectedness" and was found to be characteristic of Ss who learned about binomial probability under instruction emphasizing general concepts rather than a formula (Mayer & Greeno, in press). Thus there is some support for the claim (Gagné, 1965) that meaningful conceptual learning and the discovery and generalization of a principle result in about the same outcome.

The result of learning by rule is primarily the addition of new components to cognitive structure rather than the reorganization of existing components. These new componetns include a list of defined variables and the sequence of operations relating them. The new components may in fact have a great degree of "internal connectedness" as shown by the advantage of Ss learning by rule on Familiar problems and problems posed in the context of the formula. However, the fact that the advantage is lost when the problems require more interpretation shows that the new structural components added by rule Ss were not well integrated into existing cognitive structure. A test of long-term retention should, if this explanation is correct, show that the discovery Ss retained more information. The test of relearning after 24 hours used in the present study merely demonstrated that neither group had forgotten much instruction during that time.

A final set of conclusions concern procedures involved in studying aptitude and instructional variables. With regard to aptitude tests, a choice was obviously made in the present study for simplicity over psychometric elegance. One valid criticism is that the unreliability of the measuring instruments may have influenced the results. It is not known, for example, whether the failure of the test of arithmetic operations in the second experiment was due to allowing Ss to use calculators or the unreliability of the test. However, the degree of replication that was found between the two experiments regarding the concepts test makes this possibility less likely. The usefulness of a general ability criterion in studies of ATI is still in question. The fact that the general ability measure produced a reliable ATI on the posttest in the second but not in the first experiment suggests that its utility may be linked to the instructional material. In any case there is a tradeoff between the reliability offered by established tests of general ability, and the information concerning processes of acquisition afforded by tests specially constructed for experimental materials and instructional methods.

An unexpected result was the significant two-way interaction of ability and problem-type, and the three-way interaction of ability, problem-type and problem-context found in the analysis using the weighted average of ability test scores. Graphing these data revealed that the weighted score was more strongly related to performance on Luchins problems, particularly when posed in a formula context. Thus the weighted average of abilities was a particularly strong measure of how easily Ss could manipulate the newly learned components of the formula independently of the rule usually relating them.

Chapter I.2

Does ATI Involve Structural or Merely Quantitative Interaction?

James G. Greeno and Richard E. Mayer

The University of Michigan

Investigation of aptitude × treatment interaction (ATI) asks the question, "Who learns more with which kind of instruction?" (e.g., Bracht, 1971; Cronbach, 1967). A second question is whether qualitatively different learning outcomes are produced by different instructional procedures, that is, "What is learned with which kind of instruction?" (e.g., Roughead & Scandura, 1968). A positive answer to this second question can be obtained from certain patterns of performance on different types of posttest items, that is, from certain kinds of treatment × posttest interactions (TPI), particularly from the kind of interaction that Bracht (1971) called disordinal. If Ss having one treatment do less well on one kind of posttest item but better on another than Ss having another treatment, then the treatments probably give learning outcomes differing in structural properties, rather than merely differences in quantity of learning.

A third question combines the two about ATI's and TPI's. It is possible that Ss of different ability levels have learning outcomes with important structural differences in a single instructional treatment. This would appear in data as an aptitude × posttest interaction (API) with a disordinal pattern. If structural properties of learning outcomes depend on aptitude in different ways in different instructional treatments, we might obtain reliable three-way aptitude × treatment × posttest interaction (ATPI). Putting this in another way, when we examine API's we ask, "What is learned by which Ss?", and the ATPI is examined in relation to the question, "What is learned by which Ss with which kind of instruction?"

Earlier studies in our laboratory have dealt with the first two questions using the concept of binomial probability as material to be learned. Mayer and Greeno (in press) studied one instructional sequence that emphasized use of the formula in calculating. In a posttest using different kinds of items, significant TPI was obtained; Ss with the sequence emphasizing meanings of concepts did less well with items requiring calculation, but better on questions about the concept and on items where S had to recognize that a problem was unsolvable. We characterized the differences in terms of internal and external connectedness of acquired structures. Internal connectedness refers to connections among concepts in the structure, primarily involving arithmetic operators used in calculating. External connectedness refers to connections between concepts in the new structure and

other concepts already in S's cognitive structure, involving meaningful relations between new concepts and S's prior knowledge.

Egan and Greeno (this report) studied learning of binomial probability and probability of sequences using a discovery and a rule procedure. Significant ATI's were obtained, particularly involving aptitudes specifically relevant to the learning task. Scores on tests of specifically relevant aptitudes were related quite strongly with performance during learning and posttest performance after discovery learning, but were virtually uncorrelated with success in rule learning. In addition, there was evidence that with discovery learning Ss may have acquired structures with relatively more external connectedness while with rule learning, stronger internal connectedness was obtained. This was evidenced by a TPI in which discovery Ss performed less well than rule Ss on problems stated in terms of the variables of the formula, but better on story problems.

In the present experiment we used instructional procedures like those of Mayer and Greeno (in press), varying mainly in the sequencing of ideas and in relative emphasis on meanings of concepts and use of the binomial formula for calculation. Subjects were given pretests to allow evaluation of interaction between ability and instructional method on performance on different kinds of posttest items. The results obtained here can be compared with the pattern of ATI's and TPI's obtained by Egan and Greeno, in order to form an impression of the degree of similarity between the instructional variables used in the two studies. In addition, the pretests and posttests were designed to provide some information about API's and ATPI's and thus allow a preliminary judgment on whether structural outcomes of learning depend on S's ability.

Method

Subjects and Design

Forty-four Ss were recruited from the Human Performance Center subject pool. They were paid $1.50/hr for two sessions each lasting 1-1.5 hr. All Ss received the same material in the first session, which included training in use of the terminals, rules for forming arithmetic expressions, training in use of a calculator function and a function that returned to earlier material, and pretests.

The main experimental variation occurred in the second session when Ss received instruction in binomial probability. Four conditions were used in a 2 × 2 factorial design. One factor was the instructional sequence; the sequences differed in the order in which ideas were presented and in the emphasis given to different aspects of the material. Teaching in Sequence Form began with presentation of the binomial formula and emphasized use of the formula as an algorithm for calculation. In Sequence Form, Ss first saw the formula with a minimal amount

of interpretation. As instruction proceeded the various parts of the formula were described and S was taught how to obtain values of the three components, $C(N,R)$, $\overline{P}^R$, and $(1-P)^{(N-R)}$. Throughout instruction in Sequence Form, the amount of interpretation of variables was minimized, and emphasis was on numerical calculation.

In Sequence Gen teaching began with definitions of variables (number of trials, number of successes, probability of success) in relation to general concepts and ave explanations about how the concepts combine to form components of the formula. Sequence Gen began with individual variables and developed from parts to the whole concept. Throughout the instruction in Sequence Gen, emphasis was on explanations of the meanings of concepts, and calculation was discussed only when full conceptual explanation had been given.

The second factor varied among instructional treatments was presence or absence of test questions during training. In Condition Test, review questions were given after each of six sections of instruction, and S had to give correct answers to proceed to the next section. In Condition Self, each section ended with a statement of the form, "Now you should understand _____. If you do, go ahead. Otherwise review earlier material."

Subjects were allocated to the four instructional treatments Form Self, Form Test, Gen Self, and Gen Test on the basis of pretest scores. Up to five Ss were run at a time, and there was approximately one S per condition in each session. A Latin square design was used for allocating the Ss so that the distributions of pretest scores were approximately equal in the four conditions.

Thirty posttest questions were written to give a $2 \times 5 \times 3$ design of repeated measures. The first factor was problem format: questions were stated either in the formal variables of the equation (N,R,P) or as story problems. The second factor was problem type, with familiar problems (Fam) which were similar to problems solved in training; transformed problems (Tran) requiring an arithmetic change in data or an algebraic transformation of the formula to give a solution; Luchins problems (Luch) with direct solutions that could be hidden from S if the formula was applied thoughtlessly; unanswerable problems (Unan) which gave either impossible or insufficient i firmation for the stated problem, and questions (Ques) where S was asked about properties of the formula. The third posttest factor was problem content: some problems involved the probability of a specific sequence, some involved the combinatorial number of sequences having a specified number of successes, and some involved a binomial probability of having a number of successes in a specified number of trials. For each session of Ss the posttest problems were randomly ordered with the constraint that one question of each type occurred in each set of five problems. With this procedure, each ordering of posttest problems was given to approximately one S per condition.

Procedure and Materials

The experiment was controlled and data were recorded by an IBM 1801 computer. Subjects saw materials and entered responses through IBM 2260 display/keyboard terminals. Arithmetic expressions had to be displayed in linear form. The symbol "¬" was chosen to denote exponentiation, the symbol "#" denoted "factorial", and "*" denoted multiplication. Precedence of operations was specified as it is in FORTRAN. Several frames of instruction were used to teach S the use of parentheses to correctly form expressions like N#/(R#*(N-R)#), (for $\frac{N!}{R!(N-r)!}$).

Three pretests were given, two of which were similar to tests given by Egan and Greeno (1971). a ten-item test of probability concepts was given, including ideas like the probability of the union of disjoint events and the joint probability of intersecting events. For example one item in the test was "If 10 blue chips and 1 red chip were placed in a box, what would be the probability that you would draw a red chip?" As Egan and Greeno had found that skill in computation did not correlate with performance in CAI (presumably because S can call a calculator) an eleven-item test of arithmetic concepts like associativity and distributivity was included. Two items used were "Does (A+B)*(A+C)=A+(B*C)?" and "Write a simpler expression for (N¬2)*(N¬5)." Finally, Leskow and Smock's (1971) test where S generates permutations of the digits 1234 was included in the form used by Egan and Greeno, with S's score based on the extent to which he followed the plan of holding initial digits constant from one permutation to the next.

After the pretests, S received instruction in use of the calculator function which S called by pushing a button, and then entered an expression in numbers and operators that he wanted evaluated. Then S was instructed in use of a routine by which he could return to any of several specially numbered pages he had seen earlier. This was called by pushing a button, after which S entered the number of the page he wished to see. While looking at earlier pages, S could look ahead or back one page from the position he was in, or specify the number of another page he wanted to see, going back to the location he had come from by entering "RETURN". Practice with the page turner was given in the form of a spy problem, where numbered pages told which spies talk to each other, and these had to be reviewed to answer questions of the form "How can a message be sent from X to Y?"

The final thing done in the first session was to ask S to enter his score on the Mathematics Scholastic Aptitude Test (MSAT); 38 of the 44 Ss gave scores.

The second session was conducted 24 hr after the first. Brief reminders were given in the use of parentheses, the calculator function,

and the page turner. Then the instructional material was presented, with different Ss receiving different instructional programs. The general outlines of instruction in Conditions Form and Gen are described above and are similar to those used by Mayer and Greeno (in press). In the present experiment, each frame of instructional material had to fit in six lines of 40 characters each. Thus, instruction was divided into six sections of five to eight frames each, with each section corresponding to about a page of normal text (and to about one page of instruction given by Mayer & Greeno (in press).

In Condition Test, three to five questions followed each section such as "If success is defined as rolling an even number on a die, and we roll the die five times and success occurs three times, what is R?" If S responded incorrectly, the program branched back to earlier information or to frames that reviewed the needed information. All the test questions had to be answered correctly for S to proceed to the next section. In Condition Self, each section ended with a frame such as , "You should know what the symbol P(R|N) stands for. If you do, go ahead. If you do not, return to earlier pages."

When instruction was completed, a frame appeared telling S that he would have a test next, and that he would not be able to use the page turner during the test. S was told the calculator would be available. He was informed that some problems could not be solved because of impossible or insufficient information and that his answer to these should be "UNANSWERABLE". No time limit was imposed. Six illustrative posttest questions are given in Table 1. At the end of the posttest Ss were paid for their participation and dismissed.

Results

Aptitude × Treatment Interaction (ATI)

We first examine whether our instructional treatments interacted with Ss' aptitudes as measured by the tests we used. Analysis of each test was based on dividing Ss in each instructional condition into three groups. Scores used to form the groups and numbers of Ss in each group are given in Table 2.

Fig. 1 shows the proportion of posttest items given correctly by Ss in the various groups formed by pretest scores. The main effects of aptitude were all significant at $\alpha < .025$. Of the interactions shown in Fig. 1, those involving tests of probability concepts [$F(2,32)=2.64$, $p < .10$] and permutations [$F(2,32)=3.32$, $p < .05$] were of borderline significance. These effects involved the instructional variable of sequencing and emphasis (Form vs. Gen). The presence or absence of review tests during instruction did not have a significant

Table I.2-1

Illustrative Posttest Problems

Format	Type	Content	Problem
Formal Variables	Fam	Binomial Probability	N=4, R=3, P=.20, What is P(RIN)?
Formal Variables	Tran	Combinatorial	N=3 2, R=(2/3)*N, P=1/2. What is C(N,R)?
Formal Variables	Luch	Probability of Sequence	P(RIN)=.1138, 1-P=.28, R=2, N=5. What is P?
Story Problem	Fam	Combinatorial	A coin is flipped six times, giving a sequence of heads and tails. How many different sequences contain two heads and four tails?
Story Problem	Unan	Binomial Probability	Suppose that two people out of every nine like John Wayne movies. If a sample is taken, what is the probability that two people in the sample like John Wayne movies?
Story Problem	Ques	Probability of Sequence	Is there a difference between the probability that two dice rolled at once both come up 6 and the probability that one die rolled twice comes up 6 both times?

Table I.2-2

Scores and Numbers of Subjects in

Groups of Ss Formed by Aptitude Tests

Test	Group	Scores	Form Self	Form Test	Gen Self	Gen Test
MSAT	Low	<580	4	2	3	4
MSAT	Medium	580-700	3	5	2	2
MSAT	High	>700	2	2	5	4
Arithmetic	Low	≤5	3	3	2	4
Arithmetic	Medium	6-9	5	4	7	3
Arithmetic	High	≥10	3	4	2	4
Probability	Low	≤6	4	5	6	3
Probability	Medium	7-8	4	5	2	3
Probability	High	≥9	3	1	3	5
Permutations	Low	≤14	2	5	5	4
Permutations	Medium	15-24	5	3	3	2
Permutations	High	25-32	4	3	3	5

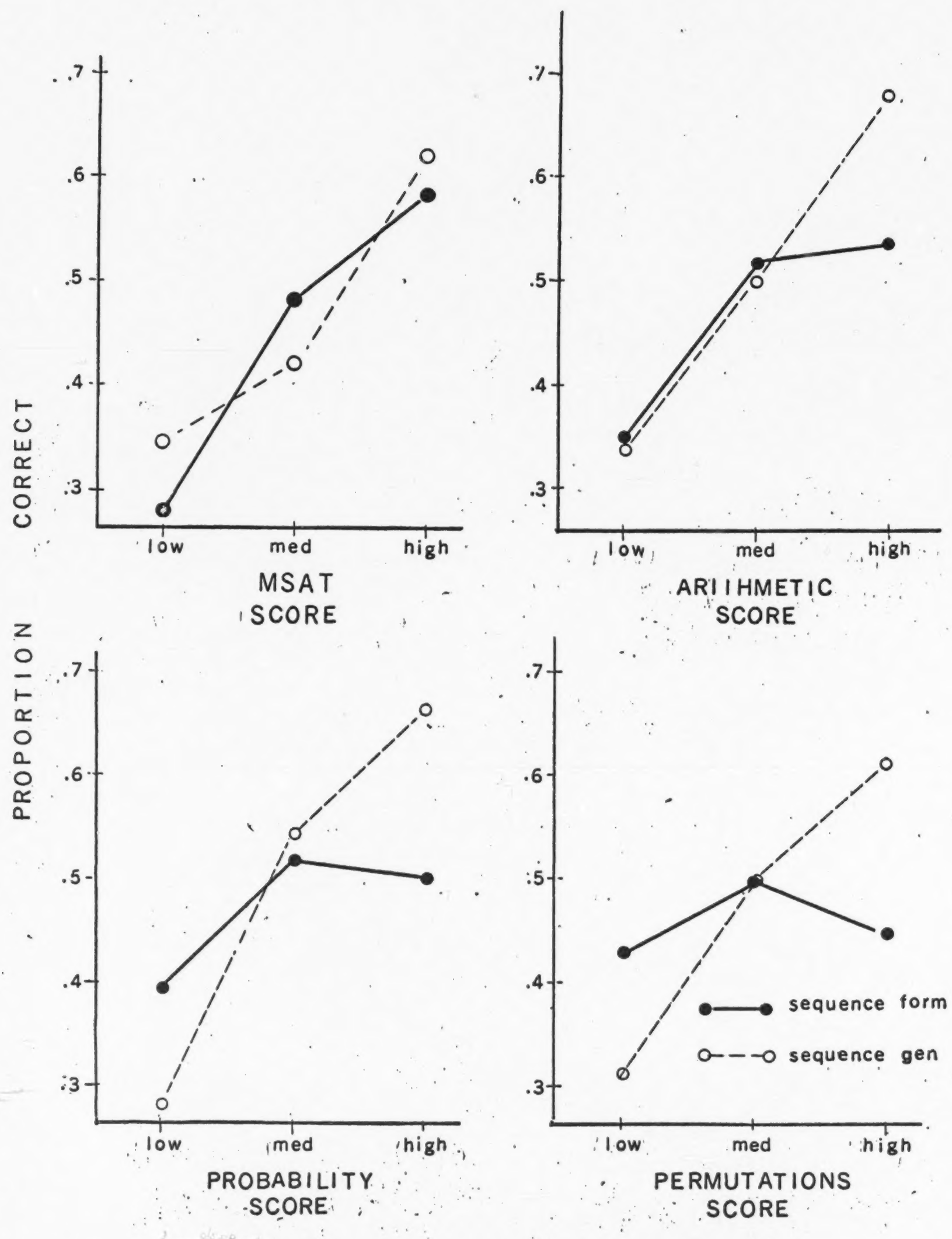


Fig. I.2-1. Aptitude x treatment interaction for four tests.

main effect nor did it interact with overall posttest performance significantly.

Treatment × Posttest Interaction (TPI)

If Ss with different instructional treatments show differing patterns of performance in the posttest, the instructional treatments may be giving learning outcomes that differ in interesting structural properties. The interesting TPI's in this experiment involved the posttest variables of format (formula vs. story problems) and content (binomial sequence probability and combinatorial problems).

First, consider the two formats in which questions were stated. Fig. 2 shows the three-way interaction involving the posttest item format and the two instructional treatment factors: sequence (Form vs. Gen) and testing (Self vs. Test). The average of the two two-way interactions (sequence × format) failed to reach significance [$F(1,40)=2.65$, $p < .15$] but the three-way interaction (sequence × testing × format) was marginally significant [$F(1,40)=3.86$, $p < .10$]. Apparently what happened was that without review questions during instruction (Condition Self) Ss with Formula sequencing acquired a structure that was more easily used in solving problems stated in terms of the formal variables used in the teaching. But Ss with instruction emphasizing the meanings of concepts acquired a structure that was more easily applied to story problems. The effect of the review questions in Condition Test seems to have been to raise performance on the kind of item for which the particular instructional sequence was weaker.

Next, consider the content of posttest questions. Fig. 3 shows the interactions of the three factors: instructional sequence, testing during instruction, and posttest content. Both of the average two-way interactions were significant; sequence × interaction [$F(2,80)=3.23$, $p < .05$] and the testing × content interaction [$F(2,80)=4.98$, $p < .01$]. The three-way interaction was not significant [$F(2,80)=1.23$, $p > .20$]. It seems clear that the Formula sequence produced superior performance on problems involving binomial probability, but inferior performance on problems involving the component concepts of sequence probability and number of combinations. The facilitating effect of review questions during instruction apparently was concentrated on the posttest items involving binomial probability.

One final interaction of interest is the sequence × format × content interaction [$F(2,80)=4.26$, $p < .025$]. The significance of this interaction shows that the differences between formula and story items of different content depended on the instructional sequence. Fig. 4 shows the three-way interaction, which turned out to involve a difference between the combinatorial items and the other two kinds of item. Compare Fig. 4 with the left side of Fig. 2. It is apparent that the interaction obtained between sequence and format was produced by the items that were about combinations.

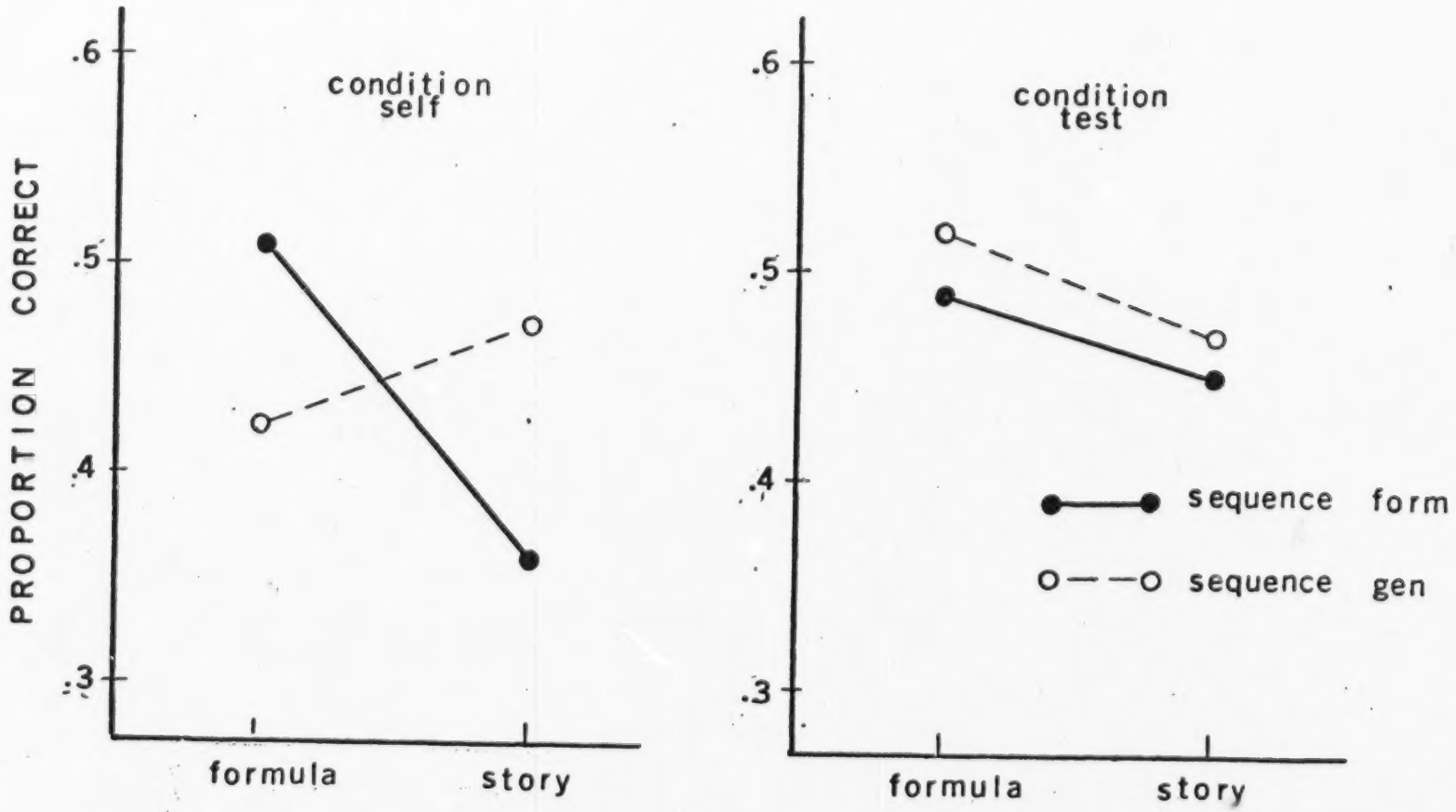


Fig. I.2-2. Interaction between instructional sequence and format of posttest items for conditions differing in use of review questions during instruction.

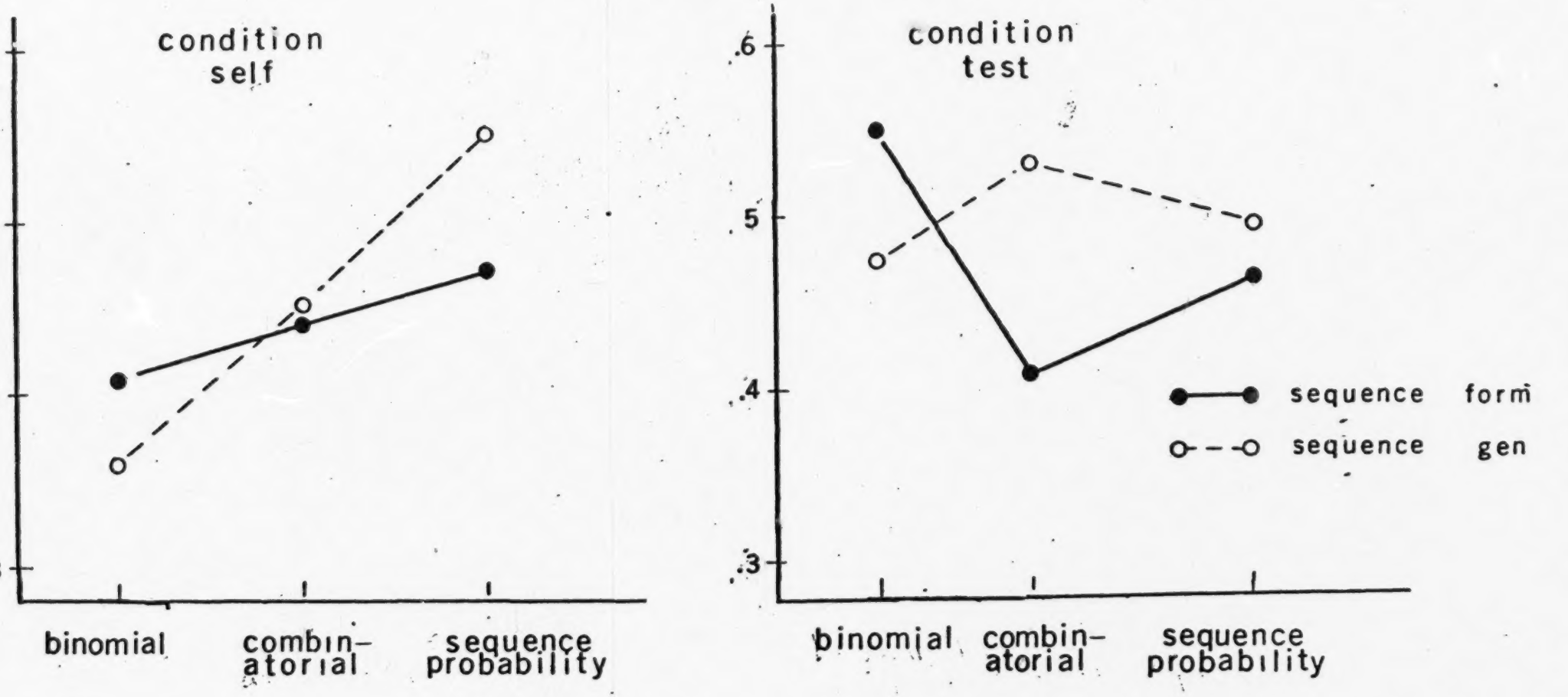


Fig. I.2-3. Interaction between instructional sequence and content of posttest items for conditions differing in use of review questions during instruction.

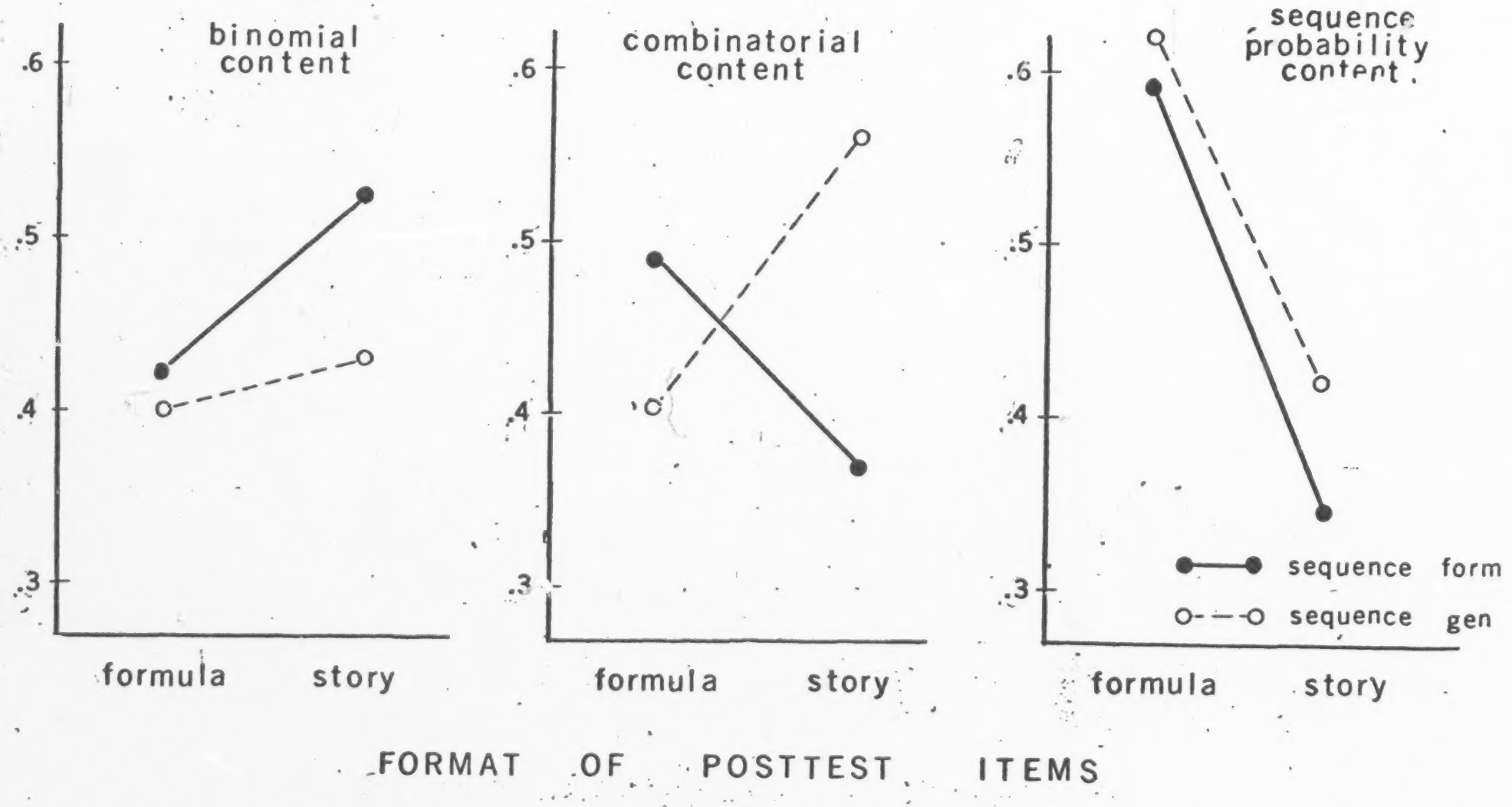


Fig. I.2-4. Interaction between instructional sequence and format of posttest items for sets of posttest items differing in content.

Unlike earlier results (Mayer & Greeno, in press) the present data did not show interaction between type of posttest item and instructional treatments. Differences between types of posttest items depended in complicated ways ontthe other factors varied in the posttest. The following interactions were all significant: type × format [$F(4,160)=19.81$, $p < .001$], type × content [$F(8,320)=2.84$, $p < .005$], type × format × content [$F(8,320)=7.36$, $p < .001$], sequence × type × content [$F(8,320)=2.83$, $p < .005$] and sequence × type × format × content [$F(8,320)=2.33$, $p < .025$]. To put it mildly, the variation of type of item did not give a homogeneous variable across the other factors. The interaction between type of item, content, and instructional sequence is graphed in Fig. 5. The relationship between sequence and type for binomial problems was like the interaction obtained in earlier data, with the Formula sequence giving better performance on familiar and transformed items, but the General-Concepts sequence giving better performance on the questions and unanswerable items. The results involving the combinatorial problems showed an interaction in the same direction as the binomial problems, although the interaction was not disordinal. But the interaction for the sequence-probability items was in the opposite direction.

Aptitude × Treatment × Posttest Interaction (ATPI)

Only a few interactions involving aptitude and treatment variables were significant, and those that were are of minor importance. One example is the interaction among score on the test of probability concepts, instructional treatment, and format of posttest items, shown in Fig. 6. The aptitude × sequence × testing × format interaction was impressively significant [$F(2.32)=11.97$, $p < .001$]. In both groups in Condition Test performance on both formula and story problems was predicted rather evenly by the test of probability concepts. In Condition Self with the Formula sequence, scores on the probability test were apparently just uncorrelated with performance on both formula and story problems. In Condition Self with the General-Concept sequence, probability test scores were correlated quite strongly with performance on both kinds of items, but with formula items medium and low ability Ss had similar performance while with story problems medium and high ability Ss all gave good performance. Another way to describe this effect involves the interaction between sequence and format shown in the left panel of Fig. 2. Apparently the better performance of Gen-Self Ss on story items than formula problems was produced by Ss of medium ability on the test of probability concepts.

Another significant APTI was the one involving probability-concept ability × sequence × type of posttest item [$F(8,128)=3.55$, $p < .001$]. The data are shown in Fig. 7. With the formula sequence, ability did not predict posttest performance very strongly, while the relationship between ability and performance was reasonably strong with Sequence Gen. However, there was a stronger correlation between ability and

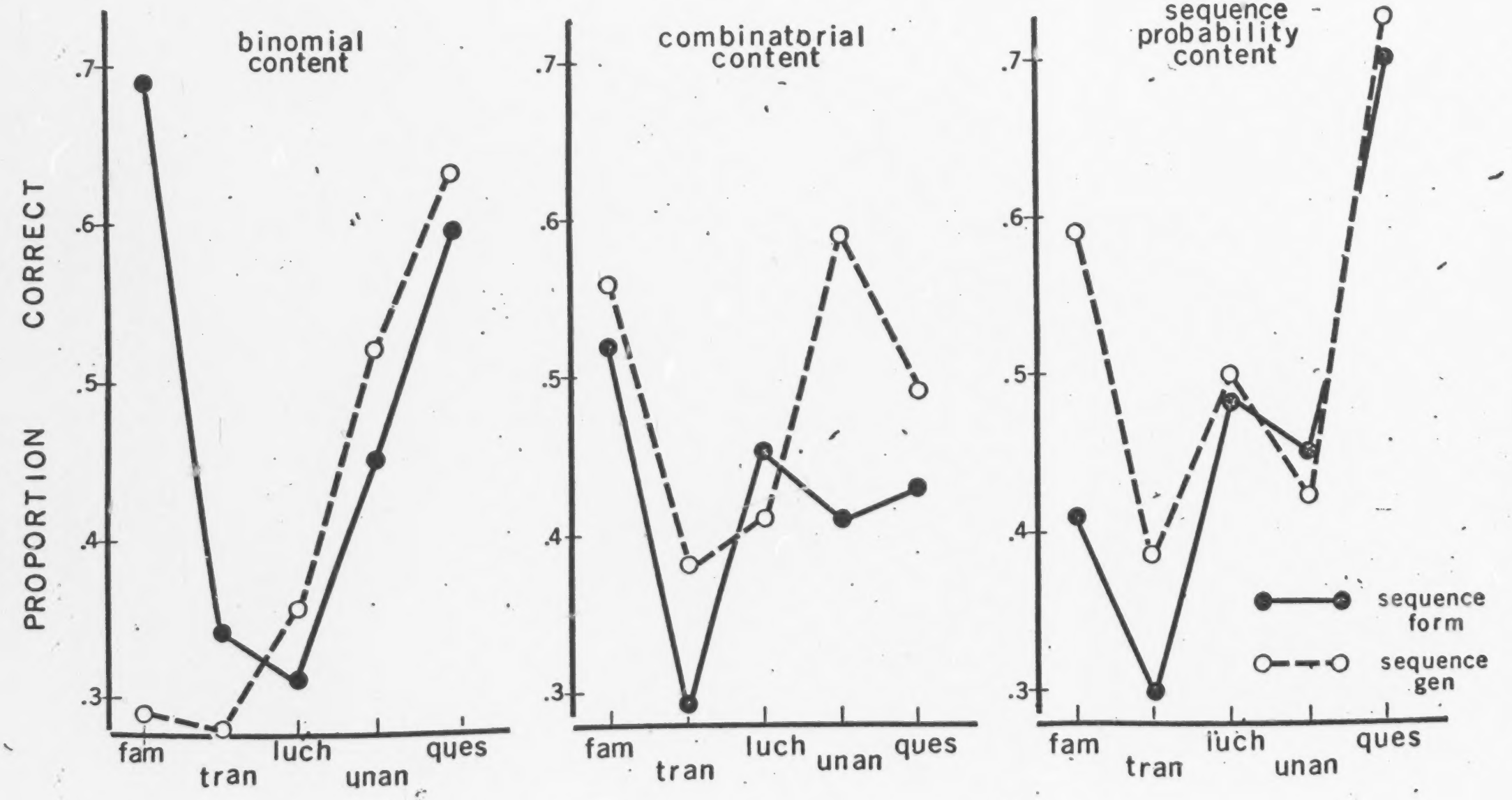


Fig. I.2-5. Interaction between instructional sequence and type of posttest items for sets of posttest items differing in content.

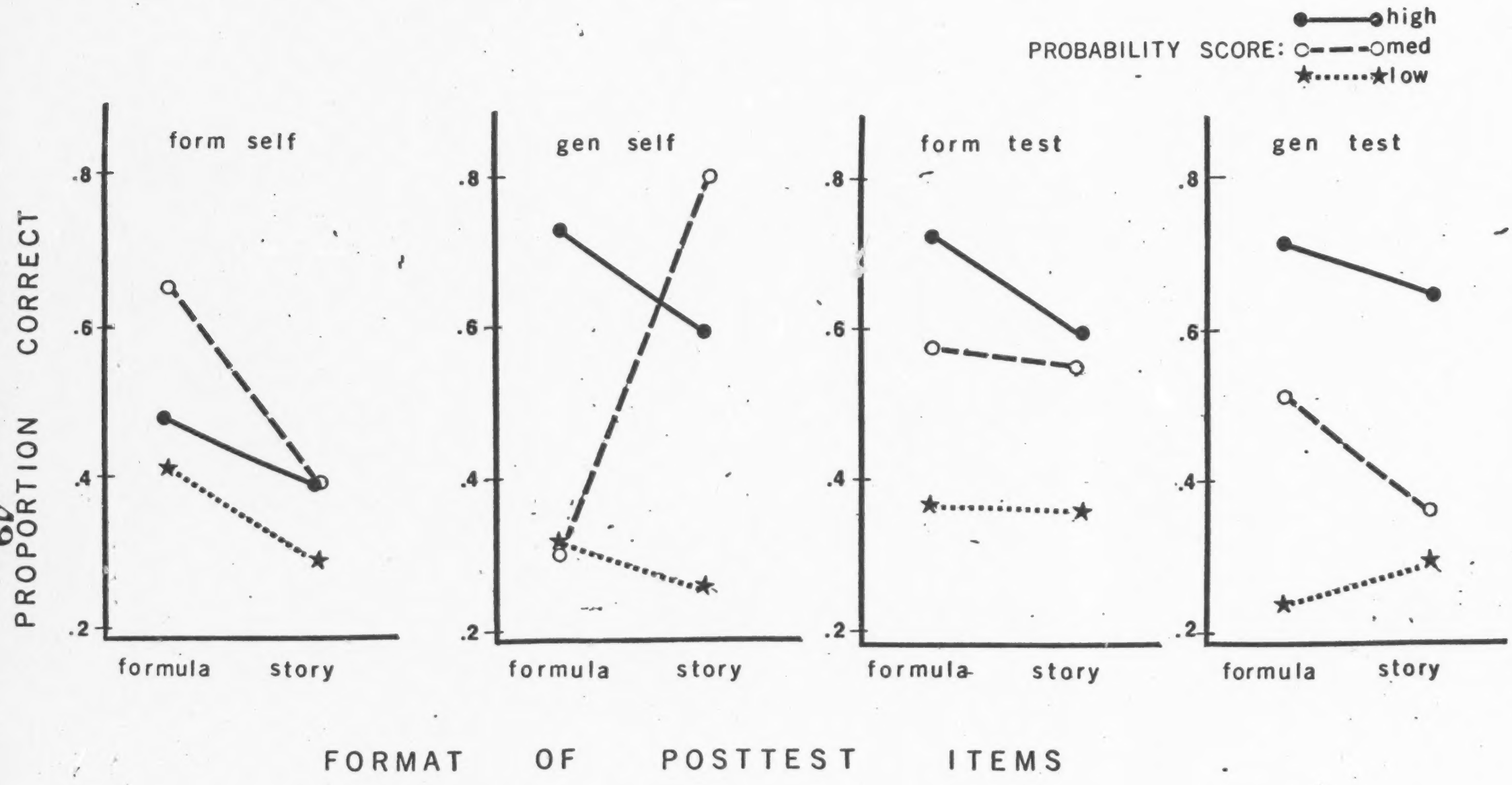


Fig. I.2-6. Interaction of familiarity with probability concepts and format of posttest items for groups of Ss differing in instructional sequence and use of review questions during instruction.

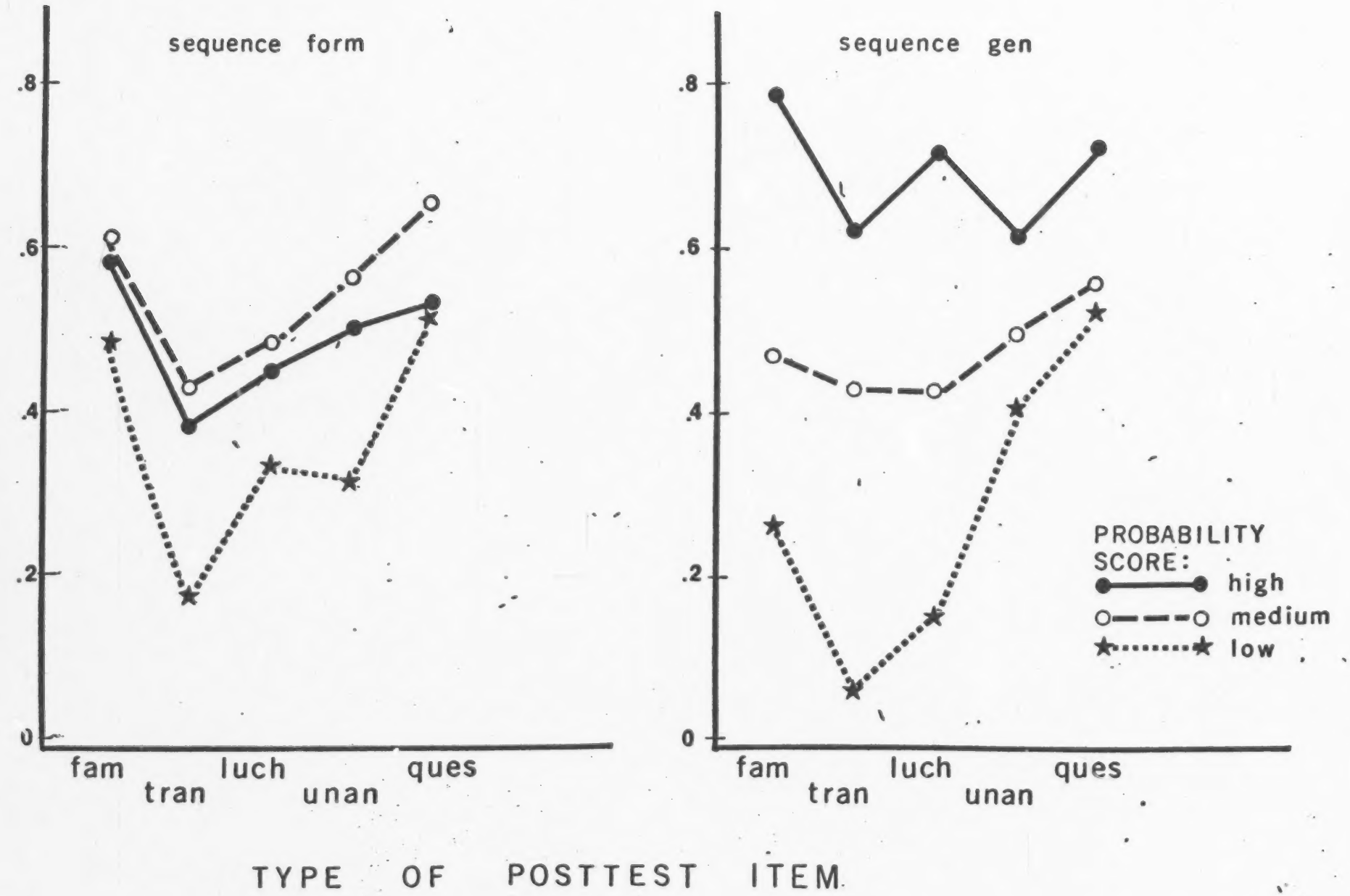


Fig. I.2-7. Interaction of familiarity with probability concepts and type of posttest items for different instructional sequences.

performance on the familiar, transformation, and Luchins problems than on unanswerable and question items. The same pattern of results was obtained with ability measured by the permutations test, where the APTI for ability × sequence × type interaction was marginally significant [$F(8,128)=1.96$, $p < .05$]. The data in Fig. 8 again show that the greater predictive power of the test under the General-Concept sequence was primarily in theffamiliar, transformation, and Luchins problems. Only one APTI was significant with each of the remaining measures, the level of significance was .05 in each case, and neither of the APTI's appeared to indicate any interesting differences in patterns of performance by Ss at different levels of ability.

Discussion

The results involving ATI's and PTI's were consistent with those obtained by Egan and Greeno (1971) and by Mayer and Greeno (in press). First, the pattern of results obtained here was so similar to that obtained by Egan and Greeno, that we tentatively conclude that the two experiments involved variables that are functionally equivalent. Egan and Greeno compared learning the binomial formula by a rule and by a discovery method while the present study, like Mayer and Greeno's compared instructional sequences that emphasized calculation a d meanings of concepts. A direct experimental comparison is needed before firm conclusions are drawn but the available results suggest that our procedures may give an example of functional equivalence between instruction that emphasizes general concepts and discovery learning.

An important aspect of the similarity between the present results and those of Egan and Greeno is the pattern of ATI's obtained. Using the test of general mathematical ability provided by the MSAT, an ATI was not obtained. Tests that predicted performance differentially in different instructional treatments were those that measured skills or familiarity with concepts that are relatively specific to the learning task. Our test of S's familiarity with concepts of probability theory has consistently provided ATI's in our experiments. The present results confirm Egan and Greeno's conclusion that measurement of S's tendency to generate a set of permutations systematically supplies information directly relevant to learning about binomial probability when S only sees his most recent response.

The test of arithmetic concepts was used in this experiment for the first time. Although the test included items involving factorials and raising fractions to powers, there was little or no evidence that abilities specifically involved in the learning of binomial probability were measured because the test of arithmetic concepts showed a pattern of results essentially like the MSAT. We now know that a test of computational skill gives substantial ATI when S has to carry out calculations. When the CAI system performs all needed calculations computational skill is uncorrelated with performance, and a more general arithmetic task provides non-differential predictions of performance.

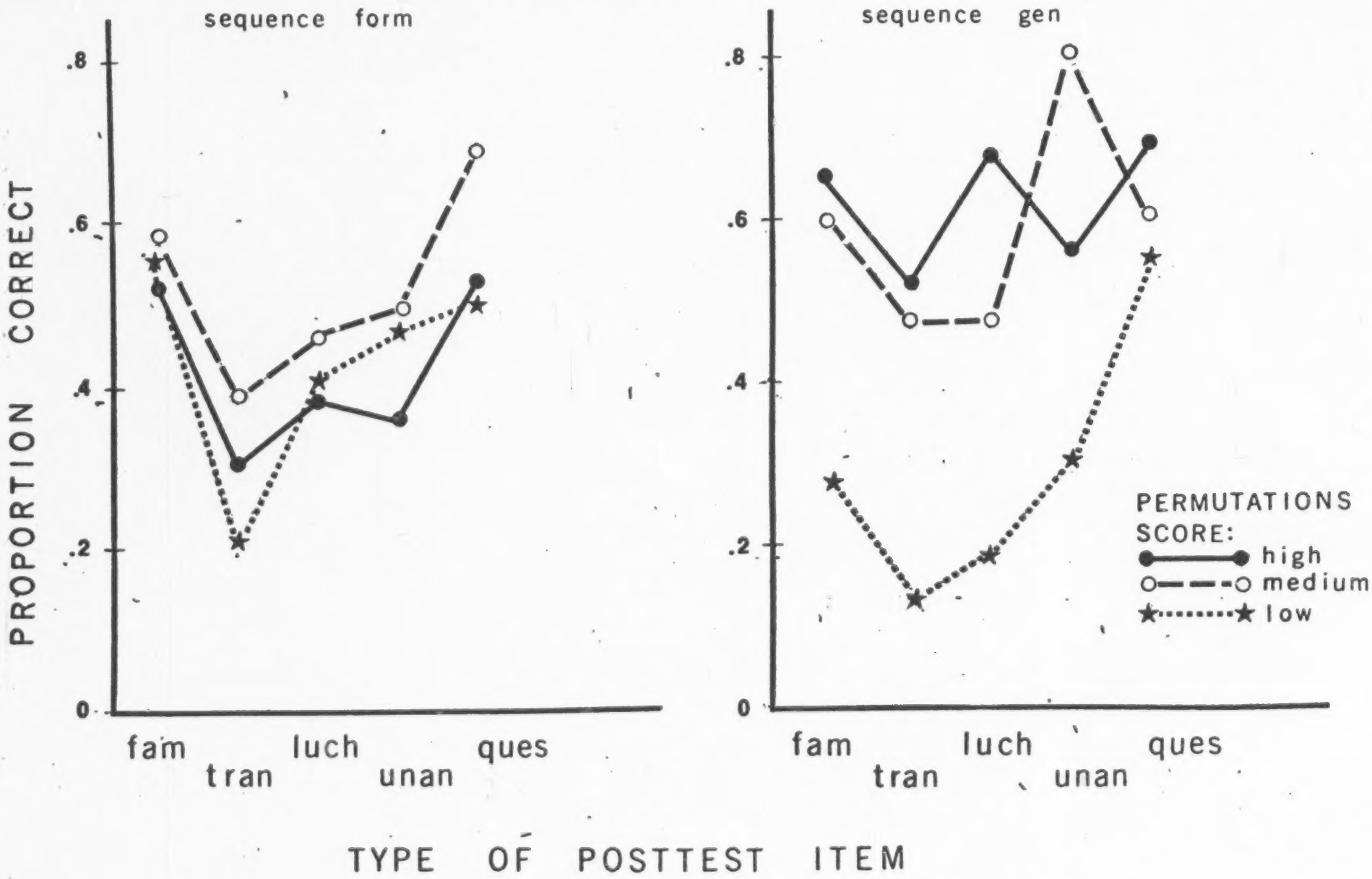


Fig. I.2-8. Interaction between use of systematic strategy in permutations test and type of posttest items for different instructional sequences.

A second dimension on which the present results corroborate Egan and Greeno's earlier findings involve the pattern of TPI's obtained. Egan and Greeno obtained a strong TPI involving the format of post-test items -- which discovery learning Ss did almost as well with story problems, as with formula problems, but with expository learning Ss did much better with formula problems as with story problems. A similar finding in the present results is that Ss with Sequence Gen did better with story problems, but Ss with Sequence Form did better with formula problems, although the effect was present only in Condition Self. This finding is not as obvious as the labels of our treatment groups might make it appear. Neither sequence described situations other than rolling a die, so the Ss with Sequence Gen were not directly trained in solving a variety of story problems. We conclude again (Mayer & Greeno, in press) that instruction emphasizing general concepts leads to a cognitive structure with stronger external connectedness and weaker internal connectedness than instruction emphasizing algorithmic calculation. Stronger external connecte ness should give a structure that can be activated in a greater variety of situations, and the better performance on story problems indicates that this was a property of the structure acquired with Sequence Gen, at least with Condition Self.

A second TPI of interest involved the content of posttest items. Subjects who had Sequence Form did better with problems involving the whole binomial concept, while Ss who had Sequence Gen did better with problems involving one of the components of the concept. Interpretation of this finding has to be tentative, but the result suggests that the structure acquired with Sequence Gen was more flexibly connected internally than that acquired with Sequence Form, in addition to its being more strongly connected externally. Solving a problem dealing with joint probability of combinations required recognition that only a component of the binomial formula was needed. The ability to use part of a cognitive structure appropriately might result simply from an adequate set of external connections giving S an appropriate understanding of the meanings of concepts involved in the structure. Another possibility is that internal connections are organized appropriately for permitting parts of the structure to be detached and used when needed. Sequence Gen was probably good for both of these characteristics since it emphasized concepts and also developed the overall structure beginning with component concepts and then gave their relationships. But in any case, the flexibility with which S can separate components of a cognitive structure probably involves an important variable.

Our present impression is that a structure with strong external connectedness and flexible internal connections is produced by instruction that either presents much information about the meanings of concepts or requires S to discover relationships in problem solving. This general impression is supported by results of several experiments, and most of the findings of these studies fit into a consistent pattern. One finding of the present experiment is mildly

inconsistent with data from Mayer and Greeno's (in press) study. Mayer and Greeno found a strong TPI involving type of posttest item, with Sequence Gen producing better performance on questions and unanswerable items but worse performance on familiar and transformation items than Sequence Form. Mayer and Greeno's result was considerably stronger in experiments using a procedure like Condition Self of the present experiment than with Condition Test. In the present data, only a weak TPI was found between instructional sequence and type of posttest item for problems involving the whole binomial concept, and items with other content failed to show this TPI (recall Fig. 5). The weak trend toward a TPI for binomial items seemed no stronger in Condition Self than in Condition Test.

Our present hypothesis is that some procedural variable may have caused the apparent reduction in the TPI of sequence × item type. One candidate involves the time given S to solve problems in the posttest; Mayer and Greeno gave 90 sec per problem, while the present experiment (like Egan and Greeno's) gave unlimited time. A second possibility involves the way in which information was presented; Mayer and Greeno's teaching booklets had a relatively large amount of information on each page, while the CAI procedure in the present experiment gave only a small amount of information on each of a larger number of frames. Further experimental work is needed to check these possibilities.

A point on which the present data provide some clarification of earlier results involves the reduction in TPI when review questions are included in instruction. Mayer and Greeno had suggestive evidence that review questions might reduce TPI's, but the comparison was across experiments using different S-populations. The present data avoid that flaw, and the strongest TPI obtained in the present experiment was virtually eliminated in Condition Test (recall Fig. 2). An interesting feature of this finding is that in each condition review questions improved performance on those posttest items on which Ss without review questions would serve to place greater emphasis on those aspects of instruction that were already strong without the reviews, thereby increasing the differences between instructional treatments. Apparently the opposite was the case. Perhaps without guidance from review questions S tends to overlook aspects of the instruction that seem not to be central to the material, and review questions produce increased attention to those ideas, thereby reducing the differences between instructional treatments.

The final issue in this experiment involves aptitude × posttest interaction (API) and aptitude × treatment × posttest interaction (ATPI). Earlier work in our laboratory has given ATI showing that Ss of different ability learn more under different instructional treatments. We also have obtained TPI indicating that cognitive outcomes with different structural properties result from different instructional treatments. A possible outcome of a study like the present one is API or ATPI indicating that Ss of different ability acquire qualitatively different cognitive structures with the same instructional treatment. No such API or ATPI was found. Of course, a single negative

finding is not decisive, but this experiment seems to have been sensitive enough to give several reliable effects and perhaps if large qualitative differences between Ss' learning outcomes occurred, they would have been detectable. Therefore, our present conclusion is that instructional treatments determine the kind of structural outcome that Ss can acquire. Further, Ss' abilities influence the amount of learning that will occur, and specifically relevant abilities may make a much larger difference in some instructional conditions than others. But we have no evidence indicating that Ss' abilities influence the kind of learning outcome that will be achieved in a given instructional treatment.

References

Ausubel, D. P., In defense of verbal learning. Educational Theory, 1961, 11, 15-25.

Bracht, G. H., Experimental factors related to aptitude-treatment interactions. Review of Educational Research, 1970, 40, 627-645.

Bruner, J. S., The act of discovery. Harvard Educational Review, 1961, 31, 21-32.

Campbell, V. N., Bypassing as a way of adapting self-instruction programs to individual differences. Journal of Educational Psychology, 1963, 54, 337-345.

Corman, B. R., The effect of varying amounts and kinds of information as guidance in problem solving. Psychological Monographs, 1957, 71, no. 2 (Whole no. 431).

Cronbach, L. J., How can instruction be adapted to individual differences? In Gagné, R. M. (ed.) Learning and Individual Differences. Columbus: Merrill, 1967.

Duncker, K. On problem solving. Psychological Monographs, 1945, 58, no. 5 (Whole no. 270).

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart & Winston, 1965.

Gagné, R. M., and Brown, L. T., Some factors in the programming of conceptual learning. Journal of Experimental Psychology, 1961, 62, 313-321.

Guthrie, J. T., Expository instruction versus a discovery method. Journal of Educational Psychology, 1967, 58, 45-49.

Katona, G. Organizing and Memorizing. New York: Columbia University Press, 1940.

Kittel, J. E., An experimental study of the effect of external direction during learning on transfer and retention of principles. Journal of Educational Psychology, 1957, 48, 391-405.

Leskow, S., and Smock, C. D. Developmental changes in problem solving strategies: Permutation. Developmental Psychology, 1970, 2, 412-422.

Luchins, A. S., Mechanization in problem solving: the effect of Einstellung. Psychological Monographs, 1942, 54, no. 6 (Whole no. 248).

Mayer, R., and Greeno, J. G. Structural differences between learning outcomes produced by different instructional procedures. Journal of Educational Psychology, in press.

Melton, A. W., Individual differences and theoretical process variables: general comments on the conference. In Gagné, R. M. (ed.) Learning and Individual Differences. Columbus: Merrill, 1967.

Piaget, J. and Inhelder, B. La genese de l'idee de hasard chez l'enfant. Paris: Presses Universitaires de France, 1951.

Polya, G. Mathematical Discovery. New York: John Wiley and Sons, 1965.

Roughead, W. G., and Scandura, J. M. What is learned in mathematical discovery. Journal of Educational Psychology, 1968, 59, 283-289.

Shulman, L. S. Psychology and mathematics education. In E. G. Begle (Ed.) Mathematics education. The sixty-ninth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1970, pp. 23-71.

Tallmadge, G. K. Relationships between training methods and learner characteristics. Journal of Educational Psychology, 1968, 59, 32-36.

Wertheimer, M. Productive thinking. New York: Harper & Row, 1959.

Wittrock, M. C. Verbal stimuli in concept formation: learning by discovery. Journal of Educational Psychology, 1963, 54, 183-190.

PART II

UTILIZATION OF COGNITIVE

STRUCTURE IN PROBLEM SOLVING AND REASONING

This part of the project report deals with an issue of considerable importance in the theory of problem solving, and thereby in the use of problem-solving tasks in education, as in discovery learning. Egan and Greeno (this report) sketched an hypothetical sequence of processes that might occur in problem solving, particularly of the kind that occurs during discovery learning. The critical step in this process is the reorganization of cognitive structure that occurs when the subject discovers a new combination of ideas or a new principle relating concepts in the situation. An understanding of such cognitive changes would form a central part of a general theory of problem solving and discovery learning.

Any knowledge that we achieve about cognitive changes during problem solving will have to be based on observation of subjects' performance--either in the problem-solving task, in related tasks presented to assess learning, or in verbal statements that subjects give about their thoughts during problem solving. Chapter I of this part reports work by John Thomas, who chose to give his subjects a problem in which overt, observable responses must be carried out sequentially in order for the problem to be solved. He used a transportation problem, where three hobbits and three orcs have to be carried across a river, using a boat that can carry two creatures, and with the constraint that hobbits can never be on a side of the river where they are outnumbered by orcs. The sequences of moves made by subjects constitute the main data of the study, and Thomas' main results involve evaluation of these sequences of overt moves as a source of information about changes in cognitive structure. His methods included quantitative analyses of the data sequences, carried out to estimate the number of significant cognitive changes occurring during problem solving and to test hypotheses that the cognitive states corresponded closely to the external states of the problem produced by the subjects' moves. He also examined dependence of performance in solving the problem on previous experience with part of the problem, and relationship between encouraging feedback given at a difficult point in the problem and performance at that point. Conclusions from these analyses were checked against the subjective reports given by subjects about the way in which they thought they proceeded in solving the problem.

Chapter II.2 gives a report of work done in trying to achieve an understanding of ways in which conclusions are drawn from premises. An understanding of ways in which subjects actually carry out deductive

inference is not only an important theoretical problem in its own right. It also is a significant problem in relation to educational practice both for understanding student performance on tests that attempt to determine whether students can extend their knowledge to situations not used directly in instruction, and in understanding the kinds of performance that can be expected of students in situations where they are left to discover the main principles involved in a lesson after concepts needed to deduce those principles have been presented.

The study carried out in this project used subjects' responses to a kind of questionnaire in which premises were presented, a possible conclusion was stated, and the subject was to indicate whether he thought the conclusion was true on the premises, false on the premises, or not decideable. One aspect of reasoning considered in the study was subjects' tendencies to assume that two objects are co-members of a class if they are co-members of another class. This tendency, known as the Von Domarus principle, was examined in several forms, including extensions to forms involving relational properties. In the items that tested the various forms of the Von Domarus principle the conclusions did not strictly follow from the premises. Between .60 and .75 of the responses were correct. But of those responses indicating that the conclusion was decideable, the number of responses agreeing with the Von Domarus principle was two to four times as great as the number concluding in the opposite direction. We conclude that a moderate tendency to follow the Von Domarus principle is probably quite pervasive in normal reasoning (an early proposal was that it characterized mainly schizophrenic thought) and that it occurs for properties of many kinds and degrees of complexity.

The other aspect of reasoning investigated was the discrepancy between ordinary intuition and the formally defined implication operator. Several hypotheses were considered, involving possibilities that subjects treat if...then in the same way as other operators such as the biconditional or the pseudoconditional. Our conclusion based on present data is that subjects have an appropriate understanding of the if...then operator, but use an additional axiom besides those in standard logic. The additional axiom is that "if p then q" implies that "if p then not-q" is false. This produces a system that is formally inconsistent, but the system coincides with most experience and in that sense it permits conclusions that are useful in ordinary reasoning.

An additional feature of ordinary reasoning shown in this study is that subjects tend not to draw conclusions that are valid but that are less specific than the premises from which they are taken. This, like the additional axiom about implication, makes the system of inference more useful in many circumstances even though it makes it inconsistent.

Chapter II.1

An Analysis of Behavior in the Hobbits - Orcs Problem

John C. Thomas, Jr.

The University of Michigan

Much attention has recently been focused on the structure of semantic information in secondary memory and the processes used to access this information. Attempts have been made to draw inferences about semantic structure on the basis of reaction time studies (Collins & Quillian, in press; Meyer, 1970). Judged similarity has also been used (Shepard and Chipman, 1970). In addition, much of the interest in free recall has centered on subjective organization (Tulving, 1966; Bower, 1970). Another possible method of studying the structure of semantic memory is through the study of problem solving. Protocols of subjects indicate that information about the world is extremely important in finding a solution, and the amount of semantic information used may be a large determinant of intersubject differences (Paige and Simon, 1966). The Gestalt psychologists held that problem solving was accomplished through reorganization similar to perceptual reorganization but at a higher level (Maier, 1930; Duncker, 1945; and Wertheimer, 1959). However, the complexity of problem solving made the early development of an adequate theory difficult for two reasons. First, there were no techniques for the on-line collection of large amounts of data in complex branching tasks, and second, there were no precise theoretical structures capable of dealing with that level of complexity. Recent advances in computer technology now allow the on-line data collection of latencies and moves for large number of subjects. Recent advances in mathematical and computer simulation models now enable us to answer precisely formulated questions about the kinds of cognitive changes that take place during problem solving.

Therefore, the experiments to be reported below had two goals. One goal was to answer specific theoretical questions concerning the relationship between external moves and changes in cognitive structure. The second goal was to investigate methods that take advantage of technological advances.

The transfer paradigm is of particular importance in discovering the relationship between learning and performance. Katona (1940) and Wertheimer (1959) both used transfer results as evidence about the generality of the information acquired as a result of problem solving or instruction. Maier (1930) found that learning the subsequences of behavior necessary to solve a problem does not necessarily produce transfer to solving the complete problem. In his study, subjects had the task of building a device consisting of three parts. Preteaching subjects to

build the parts did not facilitate solution to the complete problem in the absence of supplementary instructions. Ellis (1939) gave subjects practice on the first or second half of a finger maze and then looked for transfer to the complete maze. He found none. Tulving (1966) found similar results. He gave subjects a list of words to learn in a multi-trial free recall paradigm. After a certain number of trials, he gave the subjects another list in which the words of the previous list were embedded. During the first few trials, subjects with the practice on the part list did better than a control group who had learned a list of unrelated words. However, after the first few trials there was no evidence of transfer. In fact, the group who had the unrelated words actually did better.

One purpose of Experiment I was to determine the extent to which transfer exists between part problem and whole problem by comparing performance of subjects who had or had not solved half of the problem previously. The design allowed comparisons of state-transition probabilities and latencies as well as the gross number of moves to solution.

Although well-specified computer simulation models such as the General Problem Solver (GPS) of Newell, Shaw, and Simon (1963) have been of great value, perhaps a complementary approach would be to work on less detailed approximate models and to test individual hypotheses about the properties of human problem solving behavior. One such property is the degree to which decisions are path-dependent. It is well-known that simple choice reaction times are highly dependent on past sequences (Waugh, in press). On the other hand, the formal analysis of problems can be made purely on the basis of present position, desired position, and possible moves. Experiment I was designed to allow a test of the effects of starting point on problem solving behavior.

One simple model of problem solving is that of Restle and Davis (1962, 1963). This model assumes that problem solving may be described in terms of a Markov process, which has the following properties. The stages do not necessarily correspond to external moves made by the subject in a one-to-one mapping. The model assumes that problem solving consists of going through k stages which are equally difficult. At each instant, there is a certain probability p of going on to the next stage and a probability $1-p$ of remaining in the current stage. These probabilities are equal across subjects and stages and independent of the amount of time the subject has spent in that or any other stage. Restle and Davis (1962, 1963) presented subjects with various problems and derived theoretical distributions of solution times using parameter values estimated from the mean and variance of the solution times. In most cases, these theoretical distributions fit the data rather well; and estimates of the number of stages involved in solving the problem agreed fairly well with the subject's estimates.

Restle and Davis's model applies in situations in which the stages of problem-solving cannot be observed directly. The problem used in the present study had externally observable stages. A model with several features in common with Restle and Davis's, but adapted to the present situation, is as follows.

The probability of making the correct move from any given point remains constant independent of the number of times a subject has been in that (or any other) state and independent of how he got to that state. Furthermore, the state-transition probabilities are assumed to be the same for all subjects. The probability of a correct move at each point can be estimated from the data. The properties of this Markov model can then be tested. Experiment I was designed to allow evaluation of several aspects of these models.

Method

Subjects -- Subjects answered an advertisement for psychology subjects which appeared in the student newspaper. The majority were undergraduates at the University of Michigan. The subjects were paid $1.50 for their participation.

Procedure -- The problem was a traditional transportation problem presented in terms of hobbits and orcs. The problem is as follows. Three hobbits and three orcs were trying to cross a river. Their only means of transportation was a boat which could hold one or two creatures at a time. Every time the boat crossed the river, at least one creature had to man the boat. If at any time, even briefly, the orcs outnumbered the hobbits on either side of the river, the orcs would gang up on those hobbits and devour them. The object of the problem is to find a series of moves back and forth across the river so that all hobbits and orcs end up safely on the other side of the river. The exact instructions are detailed elsewhere (Thomas, 1971). The search graph for the problem is shown in Figure 1 and is similar to Figure 8-1 in Amarel (1968). The top line in Figure 1 gives a three number code of each state. The first digit in this code specifies the number of hobbits on the starting side of the river. The second digit specifies the number of orcs on the starting side. The third digit is 0 if the boat is on the original side and 1 if it is on the far side.

The bottom three lines in each box of Figure 1 indicate the actual display of letters given to the subject on the top three lines of the digital display used to present the problem. The word "MOVE" appeared in the middle of the screen on the fourth line. Below this, error messages were presented to the subject.

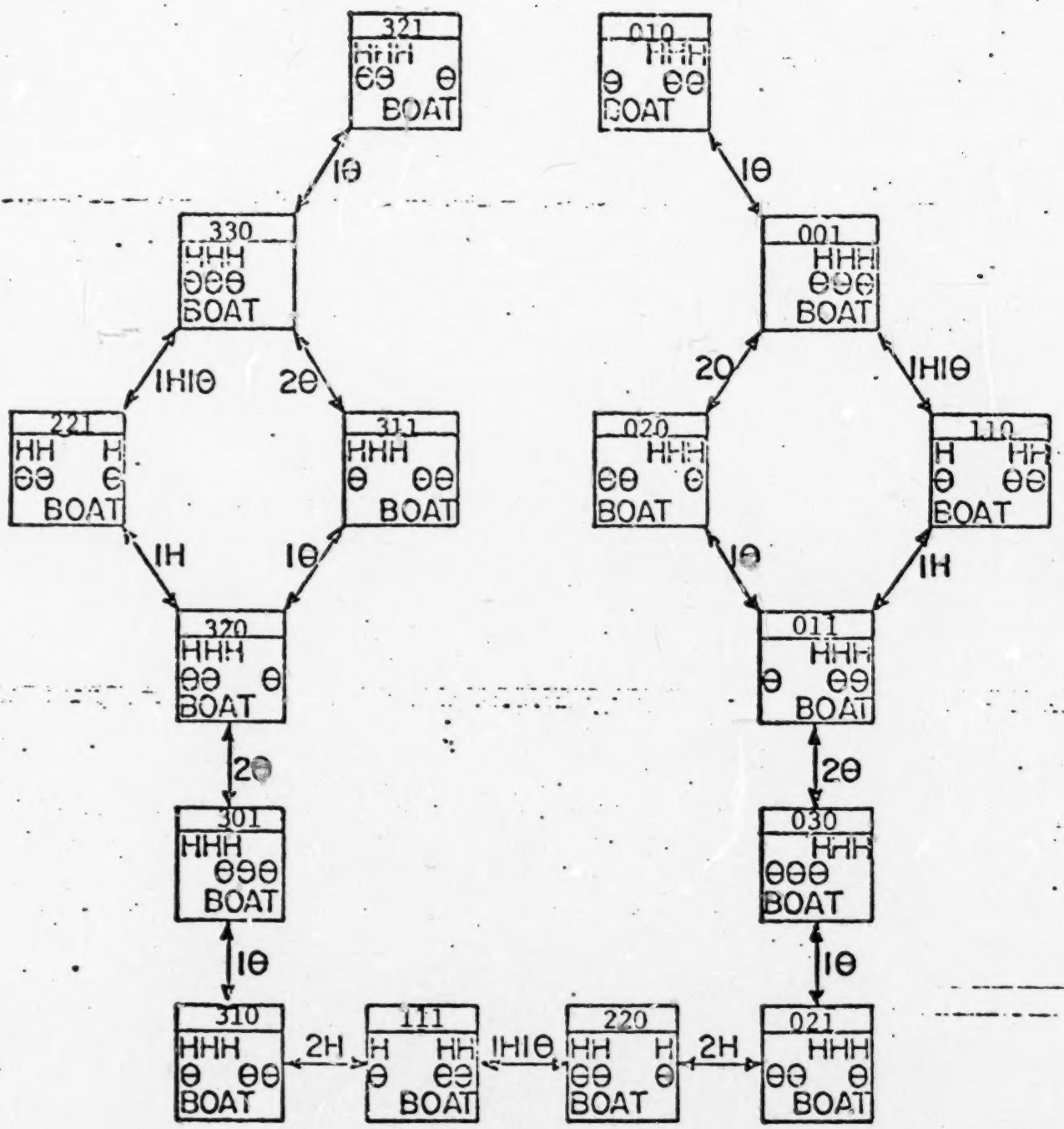


Fig. II.1-1: Search graph for the Hobbits-Orcs problem.

There were two conditions. In the control condition, subjects were given the problem in its usual version, wherein the subject begins in state 330 and works to the goal, state 001. For purposes of data analysis the problem was divided in half. All the subject's moves prior to his first entry into state 111 are considered to be in subcondition Control-First Half and all those moves made after his first entrance into state 111 are labelled as Control-Second Half.

The experimental condition was designed to assess the effects of part- learning on whole learning. Therefore, the subjects in the experimental condition were first asked to solve the second half of the problem (from state 111 to state 001). The moves the subject made during this time are referred to as Experimental-Part Problem. After solving the half problem, subjects in the experimental group were given the whole problem. Again, for purposes of data analysis, the moves prior to the first entrance to state 111 are designated as being in Experimental-First Half and all succeeding moves as being in Experimental-Second Half. It is vital to later discussions to realize that there is no formal difference whatever between the solution to Control-Second Half, Experimental-Part Problem, and Experimental-Second Half.

The subjects sat in cubicles enclosed on three sides by acoustic panel. They were given an oral presentation of the constraints of the problem and instructions on the use of the CRT display keyboard. They indicated their moves by typing a number followed by the letter H or O. Thus, moving a hobbit and an orc would be indicated by typing "1H1O".

A computer program enabled 5 Ss to run simultaneously by presenting the display representing the problem, recording each subject's move and the latency, and then computing and displaying a new representation for the state which resulted from the subject's move. If the move was in some way illegal, the program presented a diagnostic error message to the subject and left the basic display of hobbits and orcs unchanged.

Results

Transition Proportions and Latencies

Table 1 shows the overall proportion of correct moves and the average latency as a function of state. The data in this table are subject to two restrictions. First, the data include only each subject's first response in a given state. Second, only those instances in which the subject entered that state in a forward direction are included. If a subject moves backward, he may be in the process of moving further back to a still earlier state of the problem, in which case the rationale of the decision may be different. Therefore, these data were not included in the present table. However, the results were substantially the same when these restrictions on the data were

Table II.1-1

Proportion Correct and Mean Latency: Experiment I

State	Proportion	Mean Latency (secs)
330*	.798	31.668
321	.797	25.411
311	.865	16.378
320+	.342	31.336
301	.894	8.027
310	.697	30.138
111	.507	49.094
220	.753	36.272
021	.839	13.641
030	.882	8.282
011*	.908	8.469
110	.885	5.101
020	.967	6.002

Note- * denotes states where there are two correct moves;
+ denotes state where there are one correct and two incorrect moves.

eliminated (cf. Thomas, 1971). The data indicate that states 320 and 220 are moderately difficult. All the other states are relatively simple. There were a sufficient number of errors in states 330, 320, 320, and 111 other than backward moves to allow a breakdown into types of errors.

Table 2 shows the percentages of various types of incorrect but non-backward moves. As the table shows, many subjects attempted to begin the problem by moving two hobbits. In state 320, a good many subjects attempted to take over two hobbits or one hobbit and one orc. These two moves amounted to 25.2% of all moves made in state 320. In state 310 the most common response other than a backward move was to restart the problem. The second most common error was to attempt to take over one hobbit. In state 111, the most common error was to attempt to take a single orc back. Many subjects restarted at this point, and many also attempted to take back a single hobbit.

Table II.1-2

Proportions of Error Moves for Difficult States

State	1h	10	2H	20	1H10	Restart	Non-computable and other
330	.069		.552			.249	.138
320			.236		.467	.230	.067
310	.292			.083	.083	.417	.125
111	.146	.379		.049		.262	.165

Effects of Starting Point

The probability of making the correct move from state 111 was independent of whether the subject started in state 111 or reached state 111 in the course of solving the problem as shown in Table 3.

Table II.1-3

Proportions of Moves from State 111

Subcondition	Correct (1H 10)	Backward (2H)	Other, Including Restart, Illegal
1-2	.533	.233	.233
2-0	.459	.205	.336
2-2	.448	.241	.311
Total	.505	.219	.276

(In state 111, $\chi^2(2) = 2.303$, $p > .25$). A likelihood ratio test was performed to determine whether the proportion of various types of moves differed as a function of starting point. The rationale and use of this test is discussed by Wilks (1962, Chapter 13). There were significant differences between subconditions Experimental-Part Problem

and Experimental-Second Half ($\chi^2(35) = 53.384$, $p < .025$) and between Experimental-Part Problem and Control-Second Half ($\chi^2(35) = 57.173$, $p < .015$). As shown in Table 4, those subjects who get into the early states of the problem from state 111 did worse than subjects who started in state 330, went to state 111 and later backed into the early states again. This is not too surprising since subjects in Control-Second Half and Experimental-Second Half already had experience with the first part of the problem. In contrast, after state 111, the subjects in Experimental-Part Problem show no consistent advantage or disadvantage relative to the subjects in subconditions Control-Second Half and Experimental-Second Half. In states 220 and 021 the subjects in the latter two subconditions do slightly better. In state 011 the subjects in Experimental-Part Problem do better, while in states 030, 110, and 020 there are no meaningful differences.

Table 5 presents the average number of moves to solution for the various subconditions. Two-tailed significance tests were run and revealed a significant difference in number of moves between subconditions Control-Second Half and Experimental-Part Problem ($t(113) = 2.09$, $p < .05$). An F-test among subconditions Experimental-Part Problem, Experimental-Second Half, and Control-Second Half was not significant ($F(2, 156) = 1.04$, $p > .1$).

Transfer

If the major psychological process involved in problem solving is amove by move analysis based on heuristic look-ahead, any reasonable view of human memory would seem to predict large transfer from the part problem to the same portion of the game tree encountered in the whole problem. This would be manifested in a difference between the Experimental-Second Half performance and the Control-Second Half performance and also between the Experimental-Part Problem and Experimental-Second Half performance.

There was no evidence in the data for a positive transfer effect. Likelihood ratio tests did not reveal any significant differences in state-transition probabilities between subconditions Experimental-Second Half and Control-Second Half ($\chi^2(35) = 33.55$, $p > .1$). Whether one considers the number of moves to solution as independent measures or treats them as difference scores, the average number of moves to solution was not significantly different for subconditions Experimental-Part Problem and Experimental-Second Half ($t(86) = 1.241$, $p > .1$),and ($t(86) = 1.03$, $p > .1$.) respectively. And the number of moves to solution was actually greater for subcondition Experimental-Second Half than for Experimental-Part Problem. The difference in number of moves to solution between subconditions Experimental-Second Half and Control-Second Half was also nonsignificant ($t(113) = .6269$, $p > .1$).

A look at Table 4 indicates that there is no consistent superiority of either the experimental or the control condition over the other.

Table II.1-4

Proportion of Correct Moves in Various States
as a Function of Subcondition

State	Subcondition 2-0	1-2	2-2
330	.700	.868	.722
221	.526	.868	.722
311	.455	.737	.745
320	.364	.500	.403
301	.480	.670	.711
310	.395	.607	.549
111	.459	.533	.448
220	.639	.681	.741
021	.750	.865	.867
030	.911	.886	.881
011	.976	.870	.881
110	.875	.875	.889
020	1.000	.909	1.000

Table II.1-5

Average Number of Moves to Solve

Subcondition	Average Moves to Solution
Control-First Half	13.00
Experimental-First Half	10.75
Control-Second Half	15.52
Experimental-Part Problem	11.98
Experimental-Second Half	14.30

If anything, the subjects in subconditions Control-Second Half seem to have done better than the subjects of Experimental-Second Half in the most difficult states: 330, 320, 310, and 111. Even comparing Experimental-Part Problem and Experimental-Second Half we find no evidence of positive transfer at state 111. In fact, as Table 3 shows, the probability of a backwards move and the probability of other kinds of errors is greater for subcondition Experimental-Second Half than for Experimental-Part Problem.

Further evidence related to transfer is that the correlation of number of moves to solve in Experimental-Part Problem and Experimental-Second Half was a nonsignificant $r = .24$ ($t(42) = 1.65$, $p > .1$). The actual scatterplot did not indicate any curvilinear relationship which could be obscured by a linear correlation coefficient.

Another possible transfer effect is between the second half of the problem and the first. A likelihood ratio test revealed that the overall state-transition probabilities did not differ between subconditions Control-First Half and Experimental-First Half ($\chi^2(20) = 18.90$, $p > .1$). However, in terms of moves to solution, there was a superiority of subcondition Experimental-First Half over Control-First Half ($t(113) = 1.78$, $p < .05$). A look at Table 6 indicates that on the first trial, the correct response in the difficult state 320 in Experimental-First Half was more than twice as likely as in Control-First Half. The difference in proportion of correct moves on trial 1 at state 320 is highly significant ($\chi^2(1) = 11.41$, $p < .001$).

Table II.1-6

Proportion of Correct Moves in Subconditions
Control-First Half and Experimental-First Half,
First Trial Only

State	Control-First Half	Experimental-First Half
330	.808	.735
221	.788	.758
311	.792	.931
320	.219	.548
301	.889	.881
310	.861	.905

The External Markov Model

There are several assumptions of the external Markov model which can be evaluated by the data. One important feature of the model is that the state-transition probabilities should remain constant regardless of the number of times a subject is in a given state. For all subconditions and for meaningful combinations of subconditions, the differences in state-transition probabilities according to the likelihood ratio tests are highly significant. The least significant statistic was in subcondition Experimental-First Half ($\chi^2(20 = 67.90$, $p < .001$).

Table 7 shows the changes in response probabilities over trials. The states seem to fall into two classes. In one class are a group of

Table II.1-7

Changes in Response Probabilities Over Trials

State	Correct	Backwards	Other
330	Mixed	Mixed	Down
221	Down	Up	Up
311	Down	Up	Up
320	Up	Mixed	Mixed
301	Down	Up	Up
310	Mixed	Mixed	Mixed
111	Mixed	Mixed	Mixed
220	Down	Up	Up
021	Mixed	Up	Up

states in which the probability correct went down over trials, while the probability of a backward move and other errors went up. These states are 221, 311, 301, 220, and 021. The decrease in probability correct may indicate the effects of subject selection. Since these states are fairly simple to begin with, few subjects actually needed to learn the correct response.

In another class are a group of states in which the probability correct either went up monotonically over trials or showed a curvilinear trend. These are states 330, 320, 310, and 111. Apparently, these states were difficult enough so that learning at least partially compensates for the subject selection which occurred because the better Ss were in a given state fewer times. There were no statistically significant exceptions to this classification.

Subjects were in states 030, 011, 110, and 020 on multiple occasons so infrequently that reliable changes in probability correct cannot be estimated.

If the Markov model is correct, the state-transition probabilities

should also be independent of the previous state. Since the game tree for the hobbits and orcs in nearly linear, this is not a very sensitive test. This hypothesis was examined, however, for state 111. Table 8 shows the frequencies of various moves from state 111 as a function of the previous state. The row designates the state prior to entering

Table II.1-8

Frequency of 3-state Sequences at State 111

Prior State	Posterior State				
	310	220	RE	ILL	Total
310	43	139	10	34	226
220	10	6	10	6	32
start	14	29	2	2	62
TOTAL	67	174	22	57	320

state 111 and the column, the type of move made from state 111. ("Start" means the subject was starting in condition Experimental-Part Problem. RE indicates a restart; ILL, an illegal or noncomputable move.) The sequences used in this analysis do not include those in which an illegal move was made immediately prior to entering state 111. The subject's exit from state 111 did depend on his entrance ($\chi^2(6) = 47.83$, $p < .001$).

Another important assumption of the Markov model is that probabilities should be the same for all subjects. From the empirical state-transition proportions, the frequency distribution of each possible sequence can be calculated by simple multiplication, since each move is independent. This frequency distribution can be collapsed over solution sequences of equal lengths. However, the calculations are very tedious since the number of possible paths involved in solving in exactly 15 moves, for example, is very large. However, there are only 4 perfect solutions. The probability of a perfect solution was calculated to be $p = .01$, whereas the proportion in the data was $p = .095$. These results indicate rather strong individual differences producing correlations between performance at different states, rather than independence as assumed in the Markov model.

Experiment II

There were two main reasons for undertaking Experiment II. The first was to test one possible explanation for the difficulty at state 111. It was hypothesized that subjects may restart, or turn back at state 111 because they had not yet made sufficient progress. Subjects in Experiment I in subconditions Control-Second Half and Experimental-Second Half may have believed they made a wrong turn in the game tree at an earlier point. To test this explanation, some subjects in Experiment II were given feedback at state 111 that they were still on the right track. This group was designated the FB group

while the control condition which received no feedback was designated the NFB group.

The second major purpose was to measure the time needed to type and enter responses so that the mean and variance could be subtracted from the moments of the distribution of total solution times obtained in Experiment I and II to allow a more accurate test of Restle and Davis's model of problem solving.

Method

Subjects

Thirty-nine new subjects participated in Experiment II. The subjects were undergraduates at the University of Michigan who answered an advertisement for paid participants in psychology experiments. Due to a procedural error, data for eight of the subjects were lost. This left 16 subjects in the FB group and 15 in the NFG group.

Apparatus

The computer system was essentially identical to that of Experiment I.

Procedure

The problem the subjects solved was the hobbits and orcs problem, identical to that of Experiment I. However, a typing test followed the solution of the problem. In this test, a move appeared on the screen of the form used in the problem to indicate the number of hobbits and orcs to be moved (1H, 1O, 1H1O, 2H, 2O). The subject was asked to type this move on the keyboard and then shift and enter to read the message into the computer. The same random sequence of 21 moves was used for all subjects.

There were two conditions. Condition NFB was the same as the control condition of Experiment I. In condition FB subjects received a feedback message each time they reached state 111: "On the right track, solvable from here." This message was presented as part of the display at state 111, including cases in which the subject had returned to it and in which the subject had made an illegal move which left him in state 111.

Instructions for both conditions were identical to the instructions given in Experiment I, with the following modifications. Subjects were informed that they might receive a message at some point which would say "On the right track, solvable from here." This was explained to some extent. "This message means that you can solve the problem from

the point where you are. The message does not mean that your general method of going about solving the problem is good or bad; only that you are at a point form which the problem is solvable. If you do not get the message it does not mean you are on the wrong track. But if you do get the message, you are on the right track." In addition, all the subjects were told the-e would be a short test of typing speed at the end of the problem solving task.

Effects of feedback

Table 9 shows the frequencies of various moves from state 111 for conditions FB and NFB. The probability of making a correct move in

Table II.1-9

Frequence of Move Types in Feedback and No-Feedback Conditions

Condition	Correct	Backwards	Restart	Illegal	Total
Feedback	21	4	1	7	33
No feedback	21	6	5	11	43

state 111 was greater for the FB group (.637) than it was for the NFB group (.477). However, this difference was nonsignificant ($x^2(1) = 1.11$, $p > .25$).

Move Time and Typing Time

The typing test results indicated that the average typing time for subjects was 4.58 seconds/move. Analyses of variance were run for all trials combined and separately for trials 6-21 combined. If one includes all typing trials, both the mean typing time and within subject variance are rather large, whereas if one only includes the last 15 trials, the mean typing time is somewhat shorter and the within subject variance is much smaller. In neither case if the estimated mean square between subjects very large.

In general, reaction time is measured as the time interval between the presentation of some stimulus and the occurrence of a response. In Restle and Davis's (1962, 1963) model, problem solving is conceived of as a series of cognitive changes taking place over time. In their original experiments they presented problems to subjects and only measured one overt response: subjects indicated when they had solved the problem. And consequently, they only measured one time interval: the interval between the time when the subject had read the problem and the time he indicated that he knew the answer.

In the present experiments, the subject made many overt moves. What was needed then was not a measure of the interval between the onset of a particular computer display and the time the computer sensed that the subject had typed in a response. Rather, the time that was estimated was the entire length of time that the subject was thinking about a problem. For this reason, the "thinking time" is derived as a measure of the total time between the onset of the problem and the solution minus the time the subject spends typing in his moves.

It was assumed then that the total time taken for a subject to type in the $\underline{n}$ moves it took him to solve a problem was composed of three independent components: the thinking time, the average typing time times the number of moves made, and the error of measurement associated with typing in a single response times the number of moves. Let us adopt the following notation:

Let X_i = $\underline{S}_i$'s total time

y_i = $\underline{S}_i$'s average typing time

n_i = $\underline{S}_i$'s number of moves

ξ_{ij} = error of measurement of $\underline{S}_i$'s jth move

h_i = $\underline{S}_i$'s total thinking time

μ_h = the mean of the thinking times

$\mu_E = \frac{\Sigma \xi ij}{ni}$

Then $X_i = h_i + \sum_{j=1}^{n_i} (y_i + \xi ij)$

and $\mu_X = \mu_h + E(ny) + E(\sum_{j=1}^{n_i} \xi ij)$.

If $\mu_E = 0$,

then $\mu_h = \mu_X - \mu_n \mu_y$ which allows an estimate of the mean thinking time.

To obtain an estimate of the variance of thinking time, it was assumed that the error term was independent of the other components of total time and that the subject's average typing time was independent of the total thinking time and the number of moves made by the subject. With these assumptions, it is possible to derive an expression relating the theoretical variance of the thinking time to the variance of the total time and the typing time variances between $\underline{S}$s and within $\underline{S}$s from the analysis of variance of the typing test.

Let $T_i = \sum_{j=1}^{n_i} (y_i + \xi_{ij})$.

It can be shown that

$$\text{var}(T) = \text{var}(ny) + \text{var}\left(\sum_{j=1}^{n_i} \xi_{ij}\right). \quad \text{Hence,}$$

$$\sigma_T^2 + \sigma_n^2\sigma_y^2 + \mu_n^2\sigma_y^2 + \mu_y^2\sigma_n^2 + \mu_n\xi^2$$

And $\text{cov}(h,T)$ is simply $\mu_y \text{ cov}(n,h)$.

Since $\sigma_X^2 = \sigma_h^2 + \sigma_T^2 + 2 \text{ cov}(h,T)$,

then it can be shown that:

$$\sigma_h^2 = \sigma_X^2 - \sigma_n^2\sigma_y^2 - \mu_n^2\sigma_y^2 - \mu_y^2\sigma_n^2 - \mu_n\sigma_\varepsilon^2 - 2\mu_v \text{ cov}(n,h)$$

The exact value of this expression depends upon the covariance between n, the number of moves, and h, the thinking time. Unfortunately this covariance is not directly inferrable from the data. However, it seems reasonable to assume that the correlation between number of moves and thinking time is not negative. With this assumption, we can bound the variance of thinking time between two values, one of which assumes the correlation between thinking time and number of moves is 0 and the other of which assumes it is 1.0.

If $r_{n,h} = 0$, then $\sigma_h^2 = \sigma_X^2 - \sigma_n^2\sigma_y^2 - \mu_n^2\sigma_y^2 - \mu_y^2\sigma_n^2 - \mu_n\sigma_\varepsilon^2$.

If $r_{n,h} = 1$, then $\sigma_n^2 + 2\mu_h\sigma_n\sigma_h = \sigma_X^2 - \sigma_n^2\sigma_y^2 - \mu_n^2\sigma_y^2 - \mu_y^2\sigma_n^2 - \mu_n\sigma_\varepsilon^2$,

which is a quadratic equation in σ_h. Solving by the quadratic formula yields

$$\sigma_h = -\mu_y\sigma_n + \sqrt{\sigma_X^2 - \sigma_n^2\sigma_y^2 - \mu_y^2\sigma_n^2 - \mu_n^2\sigma_\varepsilon^2}$$

The upper and lower bound of the thinking time variance thus calculated is presented in Table 10. Actually, one could impose tighter strictures on the correlation between number of moves and typing time, since the covariance of total time with number of moves is equal to the covariance of thinking time with number of moves plus the covariance of typing time with number of moves.

Restle-Davis Model

According to Restle and Davis's (1962) model of problem solving, the cumulative distribution of solution times is a gamma distribution with two parameters, k and λ. One can estimate k, the number of stages

Table II.1-10

Estimated Mean and Variance of Thinking Times for Conditions of Exp. I[illegible] and II[illegible]

Exp.	Condition	Mean Thinking Time (Sec.)	Variance Upper Bound	Variance Lower Bound	Stages Upper Bound	Stages Lower Bound	Stages Ave.
I[illegible]	Control-First Half	227.5	20,109	11,729	4.41	2.57	3.49
	Control-Second Half	328.7	143,876	116,146	.93	.75	.84
	Experimental-Part Problem	245.0	67,152	49,899	1.20	.89	1.05
	Experimental-First Half	145.3	11,753	6,703	3.15	1.80	2.47
	Experimental-Second Half	183.4	64,424	45,539	.74	.52	.63
II	FB+						
	Control-First Half	219.3	19,414	15,008	3.20	2.48	2.84
	Control-Second Half	246.9	94,957	76,704	.80	.64	.72
	NFB						
	Control-First Half	164.5	9,736	6,672	4.05	2.79	3.42
	Control-Second Half	286.2	123,654	96,856	.84	.66	.75

92-76

involved in problem solving, by taking $k = E(t)^2/s^2$ where $E(t)$ is the mean thinking time and s^2 is the variance of thinking time. The time parameter λ is estimated as $\lambda = \frac{E(t)}{k}$.

Restle and Davis's model and the resultant parameter estimation of the number of stages makes several assumptions which are probably not true in detail. Since the number of stages is estimated as being $[E(t)]^2/s_t^2$, any effect which artificially inflates the variance will tend to result in an underestimation of the number of stages. One factor which will tend to cause such an artificial inflation is inter-subject differences. If the stages are not equally difficult, the estimate from $[E(t)]^2/s_t^2$ results in an underestimation of k. These two factors may help account for the consistent discrepancy Restle and Davis report (1962, 1963) between the number of stages estimated by the model and the number subjectively estimated by the subjects. There is another factor, however, which could result in an overestimation of the number of stages. If the stages are not independent, the variance is artificially decreased, since subjects who spent a long time on one stage would tend to spend less time on the next. It may also be true that only the time subjects actually spend thinking about the problem contributes to any cognitive changes. Although typing time is subtracted from the total time, it is possible that this still leaves an overestimate of actual time thinking about the problem at hand since Ss probably do not process the problem throughout the entire experimental session. Digressions may cause an overestimation in the mean thinking time and a consequent overestimate of the number of stages. The relative error in any given experiment due to the combination of these factors is difficult to estimate. For this reason, the Restle and Davis parameter estimates cannot be taken as necessarily exact. However, the ordering on number of stages within a subject population working on parts of the same problem space probably gives a good estimate of relative complexity.

With the mean and variance of thinking time estimated as outlined above, the estimated number of stages for various subconditions can be calculated. These are presented in Table 10. The estimates for the first half of the problem are consistently around .8. There are two exceptions to this rule. The estimate for subcondition Experimental-First Half ranges from 1.8 to 3.1, somewhat lower than for Control-First Half. And the estimated number of stages for subcondition Experimental-Part Problem ranges from .894 to 1.203, somewhat higher than for subconditions Control-Second Half and Experimental-Second Half. The theoretical gamma distribution along with the observed data for all subconditions Control-First Half combined is shown in Figure 2. Since conditions Control-Second Half and Experimental-Second Half did not differ significantly in any way in Experiment II, and their para-

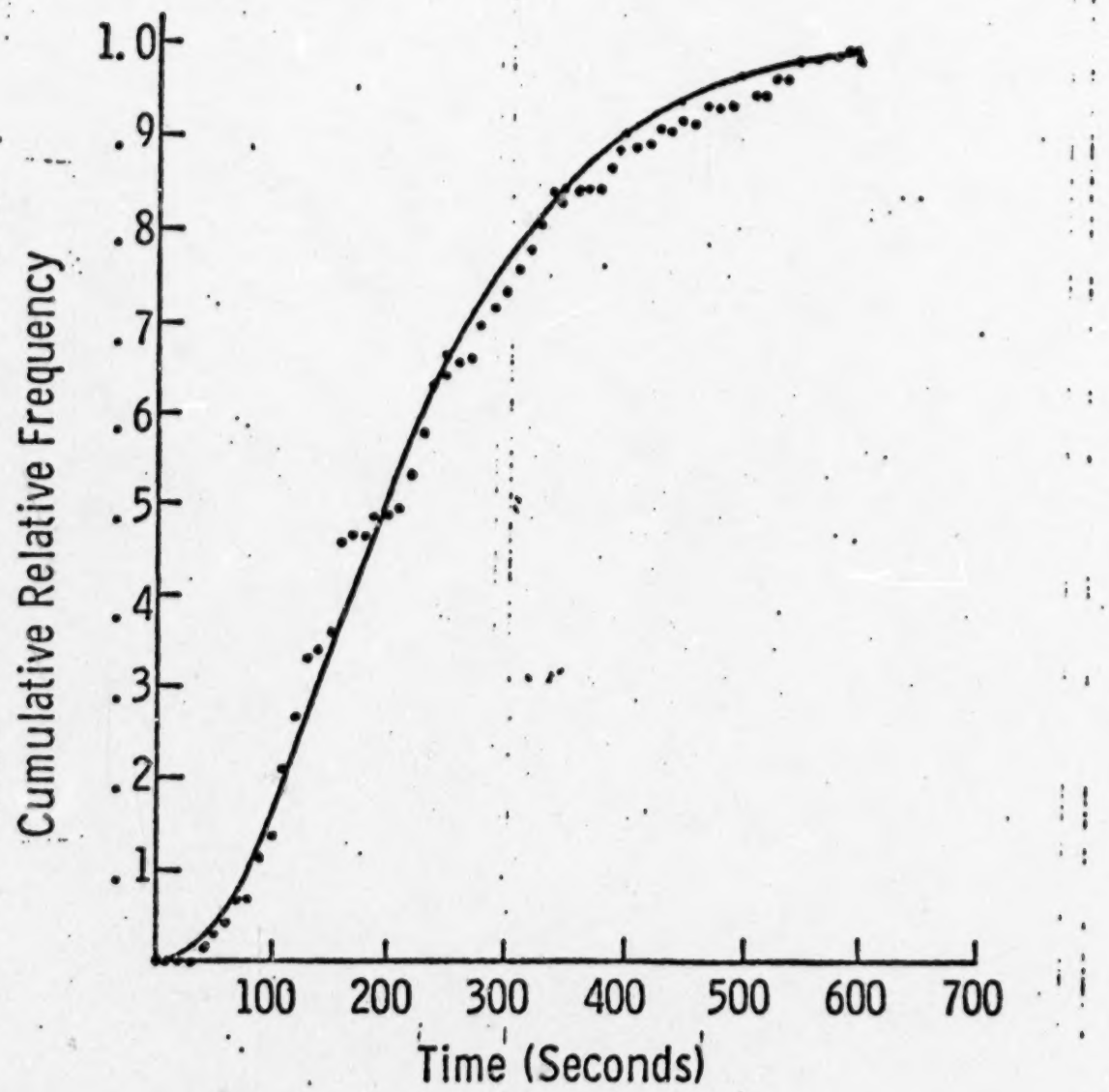


Fig. II.1-2 Cumulative relative frequency of solution time for subcondition Control-First Half of Experiments I and II combined.

meter estimates were similar, they were combined for Figure 3. None of the theoretical gamma distributions differed significantly from the cumulative distribution thus calculated.

General Discussion

The subject who solves the hobbits and orcs problem must necessarily make a series of moves or transfers of hobbits and orcs back and forth across the river. This is an inherent constraint of the problem as stated to the subjects. However, this does not mean that the primary psychological process involved in solving the hobbits and orcs problem is a move-by-move decision about which particular transfer to make next. The analysis of the problem solving-behavior of subjects in the hobbits and orcs problem may be attempted on two levels: the level of external moves or the level of internal cognitive changes. Virtually every analysis of the present results based on consideration of external moves fails, while those analyses based on cognitive change produce a degree of consistently sensible results.

A fact that may have led investigators to emphasize search processes in problem solving is that verbal protocols of subjects are filled with comments relating to "look-ahead." It may be that one's conscious awareness is highly correlated with the information held in working memory. This may be particularly true when one is asked to "think aloud." We should avoid being misled by introspective evidence into assuming that the look-ahead is the only or even the major process involved in problem solving. According to the present hypothesis, the main function of the moves that a subject makes, whether externally observable or "internal look-ahead" is to facilitate cognitive changes, where cognitive changes are conceived of as changes in the structure of the long-term semantic store of information about objects and relationships. It is assumed that a cognitive change occurs at a more general level than the learning of a new response to some stimulus situation (cf. Gagne, 1966). The present experimental results can equally well be accounted for in terms of general strategies that subjects use and the changes that take place in these strategies.

Starting Point

The results of Experiment I showed both through likelihood ratio tests and t-tests of average number of moves that a subject's behavior in a given part of an external problem solving space is not psychologically independent of how a subject reached that space. In particular, starting a problem in state 111 produced different behavior than reaching state 111 in the course of solving the problem. An analysis of the types of errors made by subjects in state 111 as a function of condition further substantiated this point. While the probabilities of moving correctly were not significantly different for subjects beginning at state 111 and those arriving at state 111 in the course of solving the

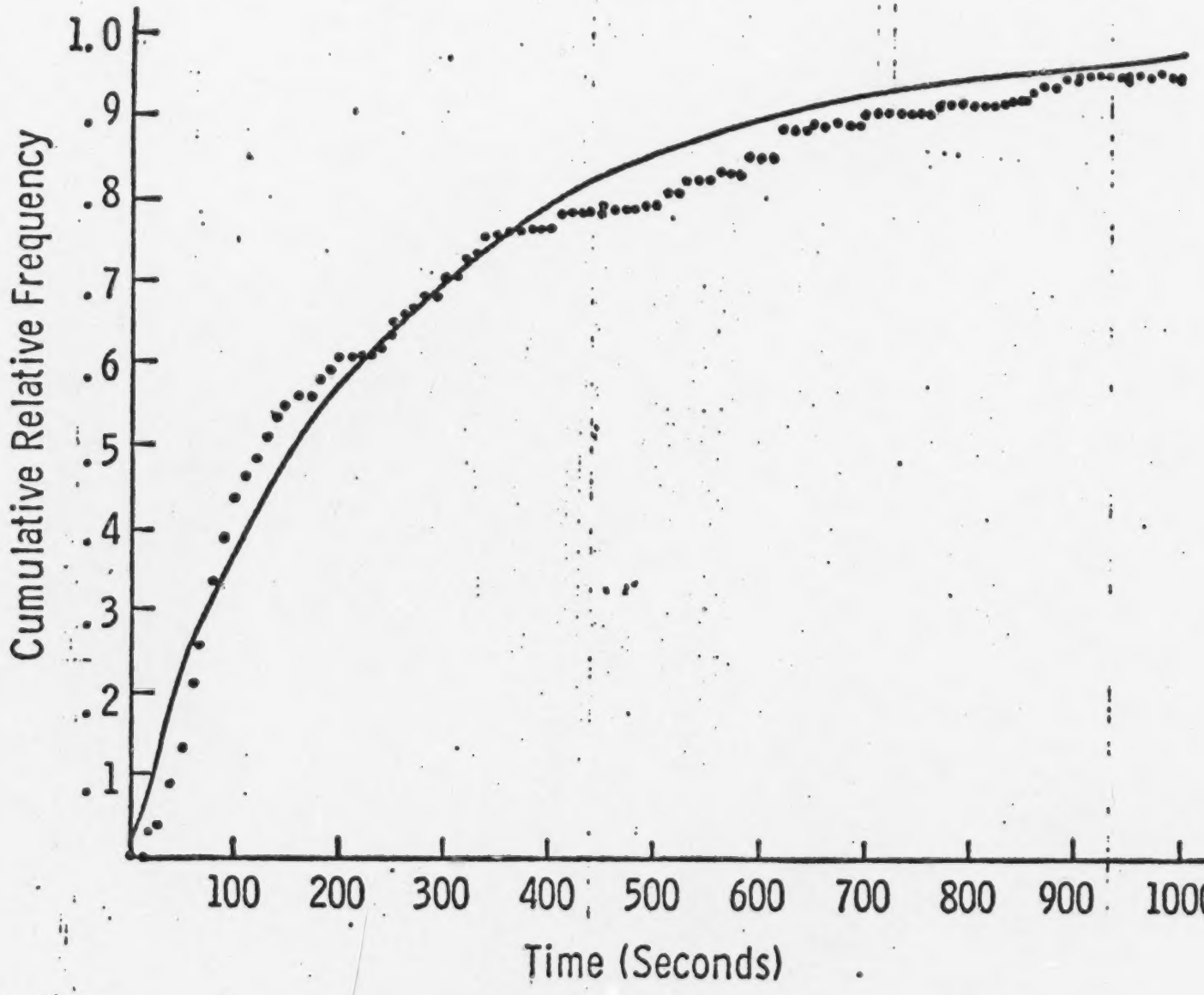


Fig. II.1-3 Cumulative relative frequencies of solution time for subconditions Experimental-Second Half and Control-Second Half for Experiments I and II combined.

whole problem, even this finding indirectly supports the difference between condition Experimental-Part Problem and conditions Control-Second Half and Experimental-Second Half since subjects starting in state 111 had two new, legal states they could move to: 310 or 220. Subjects in subconditions Control-Second Half and Experimental-Second Half moved to state 220 no more often than subjects who stated there. If we accept the most simple-minded look-ahead view of problem solving, this is a strange fact. Either subjects are unable to remember their immediately preceding state or they are unable to foresee the result of a single transformation. A chi square test, however, indicated that in fact, the move a subject made did depend on his immediately preceding state ($\chi^2(6) = 47.83$, $p < .001$).

An alternative hypothesis was evaluated in Experiment II. Perhaps a subject starting in state 111 knew he was in the right portion of the game tree. A subject arriving at State 111 in the course of solving the problem may have felt that a solution should be imminent, and, if it was not, he must have made a wrong turn and was in the wrong portion of the game tree. This hypothesis received no confirmation in Experiment II.

Experiment II also provided further evidence that condition Experimental-Part Problem differed from Control-Second Half and Experimental-Second Half in that Restle and Davis's model required different parameters for subcondition Experimental-Part Problem on the one hand, and Control-Second Half and Experimental-Second Half on the other.

Transfer

There was no evidence in Experiment I that subjects did any better on part two of a problem when it occurred in the context of a whole problem because of previous experience with that same part two presented as a separate problem. The external sequence of moves necessary to solve the hobbits and orcs problem from states 111-001 is identical regardless of starting point. However, the cognitive change necessary to see how to arrive at a solution is apparently different. Further evidence for this comes from the questionnaire results of Experiment I which showed that subjects did not necessarily recognize a position (state 111) in a problem that they had been in before when they came across that position in the middle of solving a problem, despite the fact that they could correctly recall that position later if asked to do so. These findings agree with earlier problem solving experiments by Ellis (1939) and Maier (1945), as well as with recent work in free recall by Tulving (1966).

In contrast to the lack of transfer from Experimental-Part Problem to Experimental-Second Half, there was evidence that solving the problem from state 111-001 helps subjects at state 320 when they reach it in the course of solving the entire problem. This transfer was localized in

state 320 as evidenced by the state-transition probabilities and the fact that the estimated number of stages for subcondition Experimental-First Half was one less than for subcondition Control-First Half.

Restle and Davis's model

Restle and Davis's model of problem solving applied to the data of Experiments I and II resulted in consistent estimates that the first half of the hobbits and orcs problem consisted of two or three cognitive changes, while the second half of the problem consisted of only one. These estimates agreed closely with the latency and probability correct data for Experiment I and the type of changes in probability of a correct move.

Cognitive Changes

The search space for the hobbits-orcs problem is trivial and yet people have trouble with it. If the major process in solving the hobbits-orcs problem is cognitive change as has been argued, then one might ask why are cognitive changes necessary for solving this problem and why are they moderately difficult?

To understand why the decisions necessary to solve the hobbits and orcs problem are difficult, it is necessary to remember that the problem involves the transfer of objects back and forth and that subjects have had substantial experience transferring objects in the real world prior to the experiment. However, the rules of the hobbits and orcs problem are highly unusual and present constraints seldom necessary in the real world. The particular illegal moves that subjects made in the difficult states and their retrospective reports can give us some clue as to the particular cognitive changes that are taking place in the problem. To illustrate the possible kinds of negative transfer, the difficulty at states 320 and 111 will be briefly discussed.

At state 320 the subject must move to state 301. In state 301, the hobbits and orcs are completely separated and furthermore, the orcs have the boat. This, in terms of the usual strategies one would employ for transferring or dealing with untrustworthy organisms, is absolutely unsound. The errors that subjects made at state 320, namely, moving two orcs over or moving a hobbit and an orc over, may have reflected an unwillingness to completely isolate the orcs.

At state 111 there is also a difficulty. However, this difficulty seems to be different depending on whether the subject was in subcondition Experimental-Part Problem on the one hand, or Control-Second Half or Experimental-Second Half on the other. In Experimental-Part Problem, the problem is a one stage problem of merely seeing that moving to state 220 allows the hobbits all to cross safely from which point the orcs can

ferry themselves back and forth to end the problem. Since subjects starting in Experimental-Part Problem did not have previous experience with the problem space, they may have been as yet unaware that the orcs can ferry themselves back and forth. The subject in subconditions Control-Second Half and Experimental-Second Half, on the other hand, already knew that this was possible. The evidence suggests that his difficulty was of another nature: he did not wish to undo what he had just done. The subject had to realize that a change in the proportion of the types of individuals on the far side of the river would represent progress even when the total number of individuals was constant after a sequence of two moves 310-111-220. This difficulty may be indicated by the relatively high percentage of subjects in subconditions control-Second Half and Experimental-Second Half who elected to move 1H or 10, both of which result in disaster. At least these moves represent a chance to quantitatively increase the number of individuals on each side of the river. In contrast, none of the subjects who started in state 111 chose as their first move 1 hobbits, and a smaller percentage chose 1 orc than subjects in subconditions Control-Second Half and Experimental-Second Half. The subjects in Experimental-Part Problem, since they had not themselves moved to state 111, apparently were not as concerned about the prospect of leaving the quantitative number of individuals unchanged after a single transformation. Thus, they were less likely to make these particular errors though they were less familiar with the problem solving space than subjects in Control-Second Half and Experimental-Second Half.

The transfer between subconditions Experimental-Part Problem and Experimental-First Half and the lack of transfer between subconditions Experimental-Part Problem and Experimental-Second Half both make sense in terms of these cognitive changes. For the subjects in Experimental-Part Problem, the major change was to trust the orcs to be completely isolated from the hobbits. This is precisely what needs to be done at state 320. Thus the subject who arrives at state 320 in subcondition Experimental-First Half did not need to make that particular cognitive change. In contrast, the single step needed to solve from state 111 for subjects who arrived there from state 310 is of an entirely different nature. Whether the subject was in subcondition Experimental-Second Half or Control-Second Half makes no difference. If he still did not trust the orcs, he could not have reached state 111. Isolating the hobbits and orcs is a prerequisite to reaching state 111 for these subjects. Therefore, the previous experience of the subjects in Experimental-Part Problem was of no further help to subjects in state 111 over and above what they learned in going from state 320 to 301. The move needed at state 111 for subjects in subconditions Experimental-Second Half and Control-Second Half was difficult because, in some sense, it requires undoing what they just did.

A Gestalt psychologist would not have been surprised by any of the foregoing results. The present findings seem more in line with the ideas of Katona, Maier, Kohler, Wertheimer, and Duncker than either S-R theory or the information-processing. Gestalt notions of problem solving may have failed to become more popular primarily due to an inability to demonstrate correspondences between data and theory. It is hoped that the results of the present experiments, due to advanced in data collection and theory, enable one to make at least some educated guesses about how many cognitive changes there are in the hobbits and orcs problem, where they typically occur, and what these changes might be.

In particular, the collection of latencies and particular move sequences allows one to use several kinds of analyses not readily available from a few detailed protocols. From the total solution times, one may apply Restle and Davis's model of problem solving which fit the data of several problems in the present experiments as well as that of earlier work by Restle and Davis (1962, 1963) and Davis (1964). This model allows one to estimate the number of stages involved in problem solving. From the latencies and state-transition probabilities for particular moves one may obtain information relating to where these changes take place within the problem. From a simple task analysis of the problem combined with the frequencies of relatively uncommon errors, one may gain insight into what these changes may be. By the use of such procedures it is hoped that the analysis of problem solving may be carried out on the level of cognitive changes and not on the superficial level of move choices.

Chapter II.2

Paleologic: Relationships between Human Thought and Truth Functional Logic and the Predicate Calculus

Douglas M. Stokes

The University of Michigan

It is well known that man does not infallibly follow that pattern of thinking prescribed to him by formal logic. Abelson and Rosenberg (1958) and Janis and Frick (1943), among others, have demonstrated the role of affective valence on extralogical thought, while Arieti (1955), Chapman and Chapman (1959), Dawes (1962), Gottesman and Chapman (1960), Von Domarus (1944), and Woodworth and Sells (1935) have investigated the relation of thought to Aristotelian syllogisms and simple set theory. Wason (1968), Haygood and Bourne (1965) and others have studied the use of truth functional rules in concept identification tasks. Other investigators have studied other aspects of thought. No systematic study has been done, however, to investigate the relation between the layman's deductive system and the deductive systems prescribed by truth functional logic and the predicate calculus in a natural language situation. The present study is an attempt to partially fill this vacuum.

Definitions: The signs '$\sim p$' and '$\bar{p}$' are used by logicians to denote "the negation of p" or the proposition that is true whenever p is false; '$\sim p$' is expressed verbally as "it is not the case that p" or simply "not p". The signs 'pq' and 'p&q' are used to denote the "conjunction of p and q" or the proposition that is true when both p and q are true; 'pq' is translated as 'p and q'. The following truth functions are defined in terms of conjunction and negation:

$$\text{Definition 1:}\quad p \vee q \overset{\text{def.}}{=} \sim(\bar{p}\,\&\,\bar{q})$$

$$\text{Definition 2:}\quad p \supset q \overset{\text{def.}}{=} \sim(p\,\&\,\bar{q})$$

$$\text{Definition 3:}\quad p \equiv q \overset{\text{def.}}{=} (p \supset q)\,\&\,(q \supset p)$$

Logicians verbally express '$p \vee q$' as "p or q", '$p \supset q$' as "if p, then q", "q if p" or "p only if q", and '$p \equiv q$' as "p if and only if q". The proposition expressed by '$p \vee q$' is called "the disjunction of p and q", that by '$p \supset q$' the "conditional", and that by '$p \equiv q$' the "biconditional". The sentence letter in the left position of a conditional is said to denote the "antecedent"; the right sentence letter is said to denote the "consequent".

Method

Subjects: 40 male and 40 female college students from the Human Performance Center Paid Subject Pool were used. The Ss were paid at the rate of $1.50 an hour for an hour and a half session. Seven males and seven females had received formal training in logic. These 14 Ss will be referred to as the "trained" Ss, and the rest as the "untrained" Ss.

Apparatus and Procedure: Mimeographed test booklets served as the stimulus materials. Two sets S_1 and S_2 of 150 questions each were used. These two sets of questions were combined with two different sets of instructions, I and L, to yield four different test forms ($I\&S_1$, $I\&S_2$, $L\&S_1$ and $L\&S_2$). Ten Ss of each sex took each of the question forms. It was assumed that the L instructions would set the Ss to restrict their responses to strict logical implication (or what they believed to be strict logical implication.) These Ss will henceforth be called the "logical" Ss. The L instructions read as follows:

> In each question, you will be presented with some information. Following that information, a probe sentence will appear, preceded by two asterisks. Your task is to decide whether the probe sentence is true (T), false (F) or can not be decided on the basis of the information given. For instance, consider question A:
>
> A. The dog is red.
> **The dog is not red.
>
> The probe sentence is contradicted by the information given in the question. So, in the space next to A on the answer sheet you would mark an F.
>
> B. The dog is red.
> **The dog is red.
>
> In question B, the probe sentence is implied by the information given, and so your answer would be a T.
>
> C. The dog is red.
> **George Washington died in 1957.
>
> In question C, it can not be determined whether the probe sentence is true or false on the basis of the information given, so your answer would be a 0.

The remaining Ss will be called the "intuitive" Ss; they received the I instructions, which, it was assumed, would set them to respond

more intuitively". These instructions read as follows:

> This is a test of intuitive reasoning. In each question, you will be given some information. Following that information will be a sentence preceded by two asterisks, which will be called the probe sentence. If the information given suggests to you that the probe sentence is more probably true than false (even if you do not feel that the probe sentence is absolutely and logically implied by the information given), your response on the answer sheet should be either a 1, 2, or 3 with 3 meaning that you are quite sure that the sentence is true, 2 that you are less sure and 1 indicating even less certainty. Similarly, a response of -3 would indicate that you are quite certain that the sentence is false, -2 would indicate less certainty, and -1 even less. A response of 0 would indicate that you felt the information given was completely neutral as regards whether the sentence was more probably true or false. Please answer every question, and please do not write in the test booklet.

The set of questions S_1 and S_2 were generated from the 47 skeleton question listed in the appendix. Six types of questions were used (defined by the subject matter with which they dealt): mathematical questions (M), nonsense questions (N), "color" questions, dealing with the color, size and shape of objects (C), ethnic questions dealing with characteristics of ethnic groups (E), scientific questions (S), and questions dealing with political issues (P). Each of the 47 questions appeared equally often in each question type. The question types were paired (M&N; C&S; E&P) such that, if a question of a certain type appeared in S_1, a question of the same skeleton form, but with the probe negated, of the opposite member of the type pair appeared in S_2. Thus, if skeleton question 1 appeared in S_1 as an M question of Form I ($p \supset q$, $q \supset p$), it would appear in S_2 as an N question of Form II ($p \supset q$, $**\sim(q \supset p)$). This procedure was used to counteract any T-F response bias that was present. The order of questions was given by a random permutation for both forms, and a random process decided whether the probe was to be stated in the affirmative or negative for S_1. Twelve errors were made in the typing of the test booklets, typically consisting of the omission of a "not"; these questions have been reassigned to their proper category, analyzed separately or thrown out where deemed appropriate.

The Ss were run in groups of average size twelve. Previous to this study, a pilot study was run using several smaller forms with varied instructions.

Results

Three Ss were eliminated; two, a male and a female, because they gave all positive responses to the intuitive form, which was taken as evidence of a failure to understand the instructions, and one, a female, when it was discovered that her test form (logical) had only been half completed. The appendix lists the skeleton questions as well as the remaining Ss' responses to them. The responses are given in the form of a vector (a b c d) in which the a entry is the percentage of responses which affirmed the skeleton probe (which may correspond to a "T" or "F" response, depending on the question involved), the b entry is the percentage of responses which denied the probe, the c entry is the percentage "0" responses and the d entry is the total number of responses to the question. The responses of the intuitive Ss have been scored as follows, to yield comparability with those of the logical Ss: if the response is negative, it is scored as an 'F', if 0 as a '0', if positive as a 'T'. Where two groups of responses (sex, form, question type, etc.) differed substantially in pattern from one another, they are listed separately in the appendix with the abbreviated group name to the left of the vector (i.e., MI for Male-Intuitive, F for female, etc.). The results are given in a hopefully more digestible form in the discussion section.

Discussion

A Predicate Paleologic. The first topic to be discussed is the construction of partial descriptive predicate calculus. The first systematic effort to develop such a calculus was made by Von Domarus (1944) to describe the thought processes of schizophrenic patients. He asserts that schizophrenics postulate identity of subjects on the basis of identity of predicates. Thus, claims Von Domarus, one of his patients claimed that she was the Virgin Mary because she shared with her the property of virginity. In short, the Von Domarus Principle (positive Von Domarus Principle or VDP) states that inclusion of two elements in a common class (defined by some property (predicate) will tend to suggest to S that they are in any second class together. This tendency may be expressed by the following axiom of predicate paleologic (universal quantification omitted):

$$\text{(VDP)} \qquad [F(x)\&F(y)\&G(x)]\supset G(y)$$

This axiom can be readily seen to operate in normal people in the generation of hunches or hypotheses, in dream images, and in stimulus generalization. For instance, an explorer on a distant planet, having found that all the fish on the planet with elongated blood cells are carnivorous may conclude that a bird is carnivorous on the basis of its being non-carnivorous. This is an instance of stimulus discrimination or what Arieti (1955) has termed "the Von Domarus principle in reverse" (negative Von Domarus Principle or VDN). This principle, which Arieti

claims characterizes the thinking of schizophrenics and children, asserts that the fact that two elements are not in one class together will suggest to S that they are not in any second class together. The following axiom of predicate paleologic expresses this principle:

(VDN) $(F(x)\&\overline{F}(y)\&F(x))\supset\overline{G}(y)$

Gottesman and Chapman (1960), Dawes (1962) and others have found these two axioms descriptive of normal populations as well as schizophrenic populations. In the present study VDP and VDN were tested by skeleton questions 28 and 29 respectively. An example of a C-question used to which VDP thinking would yield a response of "T" is the following:

1. Specimen X is black, as is Specimen Y. Specimen X is oval-shaped.
   **Specimen Y is oval shaped.

29.9% of the untrained Ss' responses to questions of this type were those predicted by VDP. 7.2% were in the opposite direction (such as an 'F' response to question 1), and 62.8% of the responses were '0" (N=234). For the trained Ss, these percentages were 23%, 2% and 75% respectively (N=60). The female Ss seemed to respond somewhat more randomly than the males, and the intuitive Ss, as might be expected, seemed to employ VDP more often than the logical Ss. These phenomena held throughout the data on predicate paleologic and are discussed at the conclusion of this section. For all Ss, 82.4% of the non-'0' responses were those predicted by VDP. Thus, there is evidence that some Ss employ VDP thinking some of the time.

An example of an N-question to which VDN predicts a response of 'F' is the following:

2. X is a glun. Y is not a strag. X is a strag.
   **Y is a glun.

Of all the untrained Ss responses to VDN questions, 22.4% were in the direction predicted by VDN, 4.0% were in the opposite direction, and 73.6% were '0' (N=125). For the trained Ss, these percentages were 7%, 0%, and 93% respectively (N=28). Again, there was evidence that some Ss did employ VDN thinking some of the time; 85.7% of all non-'0' responses were in the direction predicted by VDN (N=35).

It is easy to envisage an extension of VDP from 1-place to n-place predicates[1] as follows:

---

[1] In this paper, only relations and classes generated by some intuitively plausible predicate are considered, arbitrary sets of n-tuples are not. Similarly, sentences written $R(x_1,x_2,x_3)$ etc. are assumed to depend on each of the arguments of the predicate for their truth value (i.e., the relation R such that $R(x_1,x_2,x_3)$ iff. $x_1=x_2$ will not be written as above, but rather as $R(x_1,x_2)$ as the value of x_3 does not affect the truth value of $R(x_1,x_2,x_3)$).

(VDP') $(F(x_1,x_2,\ldots,x_n)\&F(y_1,y_2,\ldots,y_n)\&G(x_1,x_2,\ldots,x_n))\supset G(y_1,y_2,\ldots,y_n)$

Although it is assumed that VDP' and the axioms of predicate paleologic formulated below hold for n-place predicates in general (a pilot study indicated that they hold for at least 3-place predicates), they will be formulated in terms of 1-place and 2-place predicates as the present study has restricted its investigation to these forms. Extensions to n-place predicates are obvious th oughout. VDP' will be called Positive Relation Induction (RP) in its restricted form, which is:

(RP) $(F(x_1,x_2)\&F(y_1,y_2)\&G(x_1,x_2))\supset G(y_1,y_2)$

An example of RP thought is Bohr's model of the atom. Bohr, noticing that certain relations (such as inverse square law attraction) which held between the sun and the planets also held between the nucleus of an atom and its electrons. arbitrarily posited certain other correspondences between the solar system and the atom without any strict logical justification via his "correspondence principle".

In a like fashion, VDN may be generalized to yield the principle of Negative Relational Induction (RN):

(RN) $(F(x_1,x_2)\&\bar{F}(y_1,y_2)\&G(x_1,x_2))\supset \bar{G}(y_1,y_2)$

Skeleton questions 30 and 31 provided a test of RP and RN respectively. An M-question to which RP predicts a 'T' response was the following:

3. The function f_1 is a higher order derivative of the function g_1. g_1 is a hyperbolic function of f_1. The function f_2 is a higher order derivative of the function g_2.

   **g_2 is a hyperbolic function of f_2.

23.9% of the untrained Ss' responses to questions of this type were in the direction predicted by RP, 4.0% were in the opposite direction, and 72.1% were '0' (N=201). These percentages were 22.5%, 2.5% and 75% for the trained Ss (N=40). Again, there was evidence that some RP thinking did occur: 86.3% of all non-'0' responses were in the direction predicted by RP (N=66).

An example of an S-question to which RN predicts a 'T' response is the following:

4. Compound A has a circular molecular structure, whereas Compound B does not. Compound B does not react with Compound A. Compound X reacts with Compound Y.

   **It is not the case that Compound X has a circular molecular structure whereas Compound Y does not.

24.1% of the response of the untrained Ss went according to RN, 8.9% were in the opposite direction, and 67.1% were '0' (N=158). These percentages were 21%, 9% and 70% respectively for the trained Ss (N=33). 72.6% of all non''0' responses were in the direction of $\overline{RN}$, providing evidence that some RN thinking did occur (N=62).

RP may be generalized to form another axiom of predicate paleologic, Positive Downward Relational Induction (DRP):

$$\text{(DRP)} \quad (F(x_1,x_2)\&F(y_1,y_2)\&G(x_1))\supset G(y_1)$$

This axiom might be thought to be a special case of VDP where $S(x) = (\exists y)(F(x,y))$. As a concrete example, let F(x,y) be interpreted as "Movie X was panned by Reviewer Y" and G(x) as "Movie X is pitiful." By DRP, the information "Movie A was panned by Reviewer B" is sufficient to deduce that "Movie A is pitiful" in the absence of any knowledge about reviewer B. Pitifulness is merely seen as a prerequisite for getting panned. A 3-place predicate version of DRP might be the following:

$$\text{(DRP-3)} \quad (F(x_1,x_2,x_3)\&F(y_1,y_2,y_3)\&G(x_1,x_2))\supset G(y_1,y_2)$$

The sister principle to DRP, Negative Downward Relational Induction (DRN) is expressed as follows:

$$\text{(DRN)} \quad (F(x_1,x_2)\&\overline{F}(y_1,y_2)\&G(x_1))\supset \overline{G}(y_1)$$

Skeleton question 32 and 33 tested DRP and DRN respectively. An example of an E-question to which DRP predicts a 'T' response was:

5. John is heavier than Sam. Bill is heavier than Fred. John is a Negro.
   **Bill is a Negro.

21.2% of the untrained Ss' responses to this type of question were as predicted by DRP, 6.5% were in the opposite direction, and 72.3% were '0" (N=184). For the trained Ss, these percentages were 14%, 0%, and 86% (N=42). 78.9% of all non-'0' responses were in the direction of DRP, indicating the existence of some DRP thinking (N=57).

A P-question to which DRN predicted an 'F' response was:

6. Joe is smarter than Dick. Fred is not smarter than Henry. Joe is a Democrat.
   **Fred is a Democrat.

14.4% of the untrained Ss' responses to such questions were in accordance with DRN, 7.7% were in the opposite direction, and 80.9% were '0' (N=209). For the trained Ss, these figures were 20%, 0%, and 80% (N=51). Of all the non-'0' responses, 80% were in the direction

predicted by DRN (N=50), suggesting the existence of some DRN thought.

The following two paleological principles are logical corollaries of DRP and DRN, respectively; the first will be called Negative Upward Relational Induction (URN), and the second Positive Upward Relational Induction (URP):

(URN) $(F(x_1,x_2)\&G(x_1)\&\overline{G}(y_1)) \quad \overline{F}(y_1,y_2)$

(URP) $(F(x_1,x_2)\&G(x_1)\&G(y_1)) \quad F(y_1,y_2)$

An M-question to which URP predicted an 'F' response was the following:

7. F_1 is a hyperbolic function. G_1 is a hyperbolic function.
   F_1 is the second derivative of F_2.
   **G_1 is not the second derivative of G_2.

An E-question to which URN predicted an 'F' response was:

8. John is Polish. Tony is not Polish. Fred is smarter than John.
   **Orin is smarter than Tony.

Although the "upward" principles are logical consequences of the "downward" principles, they were not so much in evidence in the data. 11.5% of the untrained Ss responses were according to URP, 8.3% were in the opposite direction, and 81.2% were '0' (N=252). For the trained Ss, these percentages were 13.5%, 2% and 84.5%, respectively (N=52). Of the non-'0' responses 62% were in the direction of URP ($z=1.84$, $p < .05$, N=58), indicating that a small amount of URP thinking did perhaps take place. 4.8% of the untrained Ss' responses were in the direction predicted by URN, 2.4% were in the opposite direction, and 92.7% were '0' (N=126). These percentages were 7%, 0% and 93% for the trained Ss (N=28). 73% of the non-'0' responses were in the direction of URN (N=11). The reason why so little upward relational induction was found might be that the element y_2 is not mentioned in the antecedents of URP and URN and so it might appear "arbitrary" to the Ss (this element is mentioned in the antecedents of the "downward axioms).

Two trends persisted in the data regarding predicate paleologic. First, the females appeared to respond more "randomly" than the males among the untrained Ss. 19.9% of the male Ss' responses were consistent with the axioms of predicate paleologic that showed highly significant z scores (VDP, VDN, RP, RN, DRP and DRN), 3.4% were in the opposite direction, and 78.7% of the responses were '0' (N=566).[2] The

[2]These percentages are the average of the ML and MI means to control for form type between male and female Ss. The female proportion were calculated in the same manner. The intuitive and logical Ss means discussed below are similarly the average of the MI and FI means and the ML and FL means, respectively (to control for sex).

females appeared to be less cautious in the sense that they showed fewer '0' responses (65.6%), but this lack of caution did not seem to indicate any greater tendency to respond according to the above paleological axioms but to indicate a greater tendency to respond seemingly at random between 'T' and 'F'. 25.7% of the female Ss were in accordance with the paleological axioms, and 8.7% were in the opposite direction, a gain of 5.8% over the males in the first category, and a gain of 5.3% in the second (N=536). The same trend was found in the URP and URN data where, for the male and female Ss respectively, 7.8% and 10,8% of the responses were according to paleologic, 4.6% and 8.1% were in the opposite direction, and 87.6% and 81.8% were '0' (N_M=192, N_F=185).

The second trend in the data was that the untrained intuitive Ss appeared to behave according to the paleologic more often than the untrained logical Ss, without this increase being attributable to a mere rise in random responding. The logical Ss were more conservative than either the male or female group above (85.9% '0' responses), whereas the intuitive Ss were less so (58.4% '0' Responses). The logical Ss behaved similarly to the females in this respect (8.7%). However, whereas the logical Ss showed only 10.7% of their responses to be consistent with the paleologic, the intuitive Ss showed 34.8% suggesting that the intuitive instructions may have increased the amount of paleological responding as well as the amount of random responding (N_L=531, N_I=571). A similar trend held in the URN-URP data, the above percentages being, for the logical Ss, 92.4%, 6.5% and 4.9%, and, for the intuitive Ss, 77.0%, 9.4% and 13.6% (N_L=185, N_I = 192. An overall analysis of the "errors" (non-'0' responses) to predicate paleologic questions revealed a trend for significantly more "errors" to be made in the direction predicted by predicate paleologic than in the reverse direction ($t=9.94$, $df=76$, $p < .001$).

A note on question type is in order. More paleological responses were recorded to the M, N, C and S questions than to the E and P questions. This was surprising in that it had been postulated that the highest use of predicate paleologic would occur with abstract and unfamiliar materials (M and N) and with emotional or attitudinal material (E and P). It would be tampting to attribute the low scores on the affective material to evaluation apprehension (i.e., the Ss did not want to seem prejudiced or rigid), but such an interpretation is probably irresponsible in that selection of materials and subtle wording effects were probably large and uncontrolled factors.

While predicate paleologic as described above undoubtedly does characterize the way some people think some of the time, it is by no means universally used and its axioms can not be thought of as established axioms of human deduction. They probably do play a role in metaphorical thinking, dreams, hunches and related phenomena.

Table II.2-1
Percentage Responses Consistent with Predicate Paleologic, '0', and Opposite to Predicate Paleologic

Question Type	Response Type: Paleological	'0'	Opposite Direction
M&N (N=456)	29.6%	64.9%	5.5%
C&S (N=497)	24.1%	71.2%	4.6%
E&P (N=430)	13.2%	81.2%	5.6%

Causality and the Pseudoconditional. To best understand the Ss' treatment of "if-then" sentences, it seems appropriate to turn to everyday experience with causality as the basis of the understanding of the "if-then" relationship. In our interaction with the everyday world, we often observe that, in a causal relationship, if the cause is removed, its effects is removed as well. Thus, if a splinter in the hand is causing pain, removal of the splinter results in the cessation of the pain. This type of experience may lead to a conception of causality in which the cause is seen as a necessary condition for its effect (as well as a sufficient condition). Such a conception could lead to the treatment of causal relationships as though cause c and effect e were related by the biconditional formula as follows:

(1) $c \equiv e$

Even when a person is willing to concede alternative causes $c_1, c_2, \ldots, c_n$ of an event e, he still may see the occurrence of at least one of these causes as a necessary condition for the effect (as in the familiar utterance, "Everything must have a cuase."). In this latter case, the causal relation would still be viewed as a biconditional one, this time between the disjunction of the alternate causes and the effect, as follows:

(2) $(c_1 \vee c_2 \vee \ldots \vee c_n) \equiv e$

When (2) holds between some c_i and e, it will be said that the pseudoconditional $c_i \rightarrow e$ holds between c_i and e.

Definition 4: $p \rightarrow q$ iff. $(\exists r)((p \vee r) \equiv q)$

From this it follows that, if untrained people base most of their reasoning with "if-then" statements on real world causal relationships, it might be expected that they would understand by such statements (map them into) biconditional or pseudoconditional formulas. If a person maps the statement "if p, then q" into a biconditional formula, then he should be willing to deduce from this statement the further statements "if q, then p" and "if not p, then not q". Other deductions should be possible as well, but these are considered in a later section

in conjunction with another hypothesis. Skeleton questions 1 through 4 were designed to test whether Ss would make such deductions.

An example of an E-question to which biconditional mapping predicts a response of 'F' is the following:

9. If Paul is Spanish, then Ivan is Polish.
**It is not true that, if Paul is not Spanish, then Ivan is not Polish.

An example of an M-question to which biconditional mapping predicts a response of 'T' is the following:

10. If Condition 1 is satisfied, then Zavier's Theorem is provable. If Zavier's Theorem is provable, then the system L is consistent.
** If the system L is consistent, then Condition 1 is satisfied.

47.5% of the untrained Ss' responses were those predicted by biconditional mapping, 24.0% were in the opposite direction, and only 28.6% were '0', the response predicted if the "if-then" sentence is mapped into the logical conditional (N-718), suggesting that some biconditional mapping did take place. As with the predicate paleologic, the intuitive Ss appeared to be less conservative than the logical Ss (15.4% '0' responses compared with 41.5% for the logical Ss'). The intuitive Ss also have more responses consistent with biconditional mapping (56.0% compared to 38.5% for the logical Ss), while not showing as large an increase in the percent responses in the opposite direction (28.0% compared to 20.0% for the logical Ss (N_I=372, N_L=346)). A similar relationship held between the female and male Ss, who yielded 17.1% and 39.8% '0' responses, 56.9% and 38.2% "biconditional" responses and 26.0% and 22.0% "opposite" responses, respectively (N_M=370, N_F=348). There is no evidence that any biconditional mapping occurred among the trained Ss: 31% of the responses were in the direction of biconditional mapping, 31% were in the opposite direction and 38% were '0'.

Skeleton questions 6, 7 and 8 were devised to test the weaker hypothesis that "if-then" statements were mapped into the pseudoconditional formula (perhaps with the causes alternate to the antecedent implicit or undefined) and not into the logical conditional. If Ss did understand by the statement "if p, then q" the logical formula $p \supset q$, then it would be expected that from the negation of this statement, "it is not the case that, if p, then q", they would understand the formula $\sim(p \supset q)$ and that from this statement they should be willing to deduce p, $\bar{q}$ and $p \& \bar{q}$, as these sentences are all logically derivable from $\sim(p \supset q)$. If, however, by the statement "if p, then q" they understand the pseudoconditional $p \rightarrow q$ (i.e., $("?" \vee p) \equiv q$), they would not be willing to make such deductions.

[3] These means were calculated by the method of footnote 2.

The following is an M-question (derived from skeleton question 8) designed to test the hypothesis of pseudoconditional mapping against the hypothesis of conditional mapping.

11. It is not true that, if x equals 0, then y equals 3.
    **y does not equal 3.

Conditional mapping would yield a response of 'T' to this question, whereas pseudoconditional mapping would yield a response of '0'. The following is a C-question (derived from skeleton 7) to which conditional mapping predicts a response of 'F' and pseudoconditional mapping predicts a response of '0':

12. It is not true that, if the car is blue, then it is a station wagon.
    **The car is not blue.

15.5% of the responses of untrained Ss to questions derived from skeletons 7 and 8 were consistent with conditional mapping, 9.0% were in the opposite direction and 75.5% were '0', consistent with pseudoconditional mapping (N=412). 61% of the non-'0' responses were consistent with conditional mapping ($z=2.08$, $p < .05$, N=101), suggesting that some conditional mapping did take place. 10.7% of the trained Ss responses were consistent with conditional mapping, 2.4% were in the opposite direction, and 86.9% were '0' (N=84). Interestingly, all 54 of the responses of the trained Ss who received the logical instructions were '0'.

Skeleton 6 was also devised to test the hypothesis of pseudoconditional mapping against the hypothesis of conditional mapping. An N-question generated from this skeleton, to which conditional mapping would yield a response of 'F' and pseudoconditional mapping a '0', was:

13. It is not the case that, if the zorkon is relondite, then the jolon is not relondite.
    **It is not true that both the zorkon and the jolon are relondite.

This time, 59.2% of the responses of the untrained Ss were consistent with conditional mapping, 20.3% were in the opposite direction and 20.5% were '0', consistent with pseudoconditional mapping (N=409). For the trained Ss, these percentages were 48%, 8% and 43%, respectively (N=85). This result is curious, but expected, as it was obtained in the pilot study. It is curious because it indicates that the majority of Ss are willing to conclude from the sentence "it is not the case that, if p, then q" that the sentence "p and not q" is true but are not willing to conclude either "p" or "not q", indicating a certain lack of transitivity of subjective implication. A tentative explana-

tion offered when this result was encountered in the pilot study was the following: when the probe sentence implied the premise and the Ss perceived this relationship of implication, they lapsed back into intuitive, biconditional thinking and asserted that the reverse implication held as well (i.e., by biconditional thinking "if the premise, then the probe" can be deduced from "if the probe, then the premise"). Under this condition of simultaneous presentation, it was hypothesized, subjective implication becomes bidirectional. Skeleton questions 44 and 45 were included to test this hypothesis. Skeleton 44 asks whether the sentence "if p, then q" implies the sentence "p and q"; the reverse implication holds whether "if p, then q" is mapped into the conditional, biconditional or pseudoconditional. An example of an S-question generated from skeleton 44 to which this type of bidirectional implication predicts a response of 'F' was the following:

14. If the specimen has traces of radiation poisoning, then it is likely that is a mutant.
**It is not both true that the specimen has traces of radiation poisoning and that it is likely that it is a mutant.

70.5% of the responses of the untrained Ss to this type of question were consistent with the hypothesis of bidirectional implication, 6.8% were in the opposite direction, and 22.6% were '0' (N=190). For the trained Ss, these percentages were 52%, 7% and 40%, respectively (N=42). This finding contrasts with that obtained with skeleton question 45. Skeleton 45 has "if p, then q" for its premise as did skeleton 44, however, it probe sentence was merely "q" rather than "p and q". q does not imply $p \equiv q$ or $p \rightarrow q$ nor does it psychologically imply "if p, then q", as will be seen in the section on information reducing deductions, below. In this case, the bidirectional implication effect should not be found just as it was not with skeletons 7 and 8. An N-question generated from skeleton 45 was the following:

15. If glacks whool, then calks perundulate.
**Calks perundulate.

12.7% of the untrained Ss affirmed the conclusion "q" to this type of question, 13.8% denied "q", and 73.5% responded '0' (N=189). For the trained Ss, these figures were 3%, 3% and 94% (N=33). Again lack of transitivity is obtained: "if p, then q" subjectively implies "p and q" but not "q". It would be interesting to determine over what types of implication this bidirectionality holds. For instance, although "p and q" implies "p", it would be most surprising to find bidirectionality in this case (i.e., "p" implying "p and q"), although no data are available. It is also apparent from the lack of transitivity observed that the two terms must be presented to the S before bidirectionality occurs. For some reason, the Ss will not spontaneously evoke for a sentence "p" a subset of the sentences q_i which imply p, allowing bidirectionality and chaining to occur, resulting in sentences derivable from the q_i now being derivable from p.

The Extra Axiom of Truth Functional Paleologic. It has been assumed that untrained Ss base much of their reasoning with "if-then" sentences upon their experiences with causal relations in the real world. In general, it would be thought that such Ss are not interested in conditions that contradict known facts about the real (or hypothetically real) world. For instance, they would not be interested in what would happen if 1 were not equal to 1, because this condition is impossible in any conceivable world. In so far as the world is logically consistent, conditions that are consistent with the real world can not have effects that are either self-contradictory or contradict the real world. It is here hypothesized that Ss generalize from this experience and will not tolerate two "if-then" statements with the same antecedent and logically contradictory consequents. If this analysis is correct, the following would be an axiom of truth functional paleologic:

(A1) $(p \rightarrow q) \supset \sim (p \rightarrow \bar{q})$

By A1, the following two sentences would be subjectively inconsistent, although they are logically inconsistent, even when mapped into the biconditional:

(a) If either Theorem 1 or Theorem 2 is true, then x is equal to y.

(b) If Theorem 1 is true, then x is not equal to y.

The above two sentences are consistent only if Theorem 1 is false (in which case the condition contradicts the real world). Skeleton question 5 was designed to test whether the sentence "if p, then q" would be a sufficient condition for the Ss to deduce that the sentence "if p, then not q" was false. This result would follow if the Ss either (1) employed A1 is reasoning or (2) mapped both sentences into the biconditional [as 'p≡q' implies ' $(p \equiv \bar{q})$]. An example of a C-question generated form skeleton 5 to which both biconditional mapping and A1 predict ab 'F' response was the following:

16. If the stone is round, then it is white.
**If the stone is round, then it is not white.

85.8% of the untrained Ss' responses to this type of question were consistent with A1 and biconditional mapping, 10.1% were in the opposite direction, and 4.0% were '0' (N=375). For the trained Ss, these percentages were 95.2%, 3.6% and 1.2%, respectively (N=85). The largeness of the proportion of responses consistent with these two hypotheses to this type of question compared with the proportion of responses consistent with biconditional mapping to skeleton question 1 through 4 suggests the probable influence of A1. A test of A1 independently of biconditional mapping was provided by skeleton questions 9 and 10. An example of a C-question generated form skeleton 9 to which A1 predicts a response of 'T' and biconditional mapping predicts a response of '0' was the following:

17. If the marble is either blue or red, then we are sampling from the third urn.
**It is not true that, if the marble is red, then we are not sampling from the third urn.

84.4% of the responses of the untrained Ss were consistent with A1, 8.9% were in the opposite direction and 6.7% were '0' (N=315). For the trained Ss, 94% of the responses were consistent with A1, 4.5% were in the opposite direction and 1.5% were '0' (N=66). In so far as these proportions were in the range obtained for simple logical deduction such as modus ponens and modus tolens (see appendix), it must be assumed that the Ss accept A1 as an axiom of truth functional logic. But the result of adding A1 to the axioms of truth functional logic results in a paleologic which corresponds to an inconsistent axiom system.

The Inconsistency of Truth Functional Paleologic. The paleologic which results from adding A1 to the laws of truth functional logic is inconsistent - as is seen in the following deduction of a contradiction in this system:

Suppose there exists a true sentence "p" under the paleologic. The- the sentence "p or q" is true, and hence "if not p, then q" is true, which implies that the sentence "if not p, then not q" is false, by A1. Similarly, "p or not q" is true, hence "if not p, then not q" is true and we have a contradiction.

Thus, under the paleologic, no sentence is true. In particular, for any sentence "r", both "r" and "not r" are false. But this is impossible in that the disjunction of "r" and "not r" is a theorem of truth functional logic and hence of the paleologic.

The reason people are unaware of this inconsistency is that they are reluctant to make the information reducing and hence counterintuitive deduction of "p or q" or "if not p, then q" from "p" (see next section for discussion); they do however recognize these deductions as valid (see next section).

(The fact of truth functional paleologic's inconsistency refutes the assertion of J. R. Lucas (1964) and others that man is fundamentally different from machines because man can assert his own consistency (or recognize his own Gödelian formula as true) while remaining consistent. The present study indicates that man is only able to do these things because he is inconsistent, as Gödel's theorem states.)

Information Reduction and Intuitiveness of Deductive Steps. In the data, it was found that, whereas Ss were willing to deduce disjunctions from conditionals and vice versa when such deductions were logically valid, they were unwilling to deduce disjunctions and conditionals (or

biconditionals) from atomic sentences or conjunctions thereof when these deductions were logically valid. The determining factor in whether an S would make a deductive step seemed to be whether the step was information generating or preserving or information reducing. It would not be surprising to find that, in ordinary thinking, Ss would be reluctant to go from a position of knowledge as to the truth or falsity of a sentence to a position of lack of knowledge as to its truth value and that they in fact prefer the reverse direction. For this reason, the logical step from the sentence "p" to the disjunction of p and some second sentence q might be counterintuitive as an isolated deduction (as it involves a loss of knowledge as to the truth value of p), whereas it might be intuitively acceptable as a substep in a larger deduction designed to go from a position of ignorance as to the truth value of a sentence to a position of knowledge of that truth value. For the same reason, the deduction of the pseudoconditional $p \rightarrow q$ from the sentence q or the conjunction p&q might also be counterintuitive. Similarly, the deduction of the pseudoconditional $\bar{p} \rightarrow q$ might be counterintuitive not only in reducing knowledge, but in involving the use of a counterfactual "if-then" sentence, which has been hypothesized to be alien to the habits of untrained Ss. However, the step from the disjunction p q to the "if-then" statement "if p, then q" would not be counterintuitive as they involve neither a reduction in information not the use of a counterfactual "if-then" statement. Skeleton questions 17 through 26 provided a test of these assumptions. Skeletons 17 through 22 required information reducing deductions; of these, 20 through 22 provided for exclusive "ors" and biconditional (or pseudoconditional) "if-then" statements, whereas 17 through 19 did not.

An example of an s-question generated from skeleton 17, to which logic predicts a response of "F" was:

18. Substance X is titanium.
    **It is not true that Substance X is either titanium or einsteinium.

50.2% of the untrained Ss gave the response predicted by logic to this type of question, 36.7% gave the "opposite" response and 13.0% responded '0'. There was evidence of some logical thinking: 57.8% of the non-'0' responses were in the direction of logic ($Z=1.97$, $p < .05$, $N=161$).

Skeleton 18 generated the following N-question, to which logic demands a response of 'T'.

19. Qualks lenerate.
    **If sants remur, qualks lenerate.

27.4% of the untrained Ss responses to this type of question were logical, 12.1% were in the opposite direction and 60.5% were '0' (N=248). The evidence for some logical thinking was nonsignificant: 69.4% of the non-'0' responses were in accordance with logic ($z=1.55$, $p > .05$, N=98).

Skeleton 19 generated the following M-question, to which an 'F' response is consistent with logic:

20. x equals 3.
    **It is not true that, if x does not equal 3, then y equals 3.

6.4% of the responses of the untrained Ss to this type of question were logical, 10.9% were in the opposite direction and 82.7% were '0' (N=156). Perhaps, the low proportion of logical responses to this type of question is due to the fact that the probe is a counterfactual "if-then" sentence. Of all the untrained Ss responses to skeletons 17-19, the questions which did not allow for exclusive "ors" or bi-conditional mapping, 29.0% were logical, 19.5% were in the opposite direction and 51.4% were '0' (N=589). These figures were 41.6%, 7.2% and 51.1% for the trained Ss (N=137).

The following skeleton questions required information reducing deductions that were valid for exclusive "ors" and biconditional mapping; the logical responses is affirmation of the skeleton probe for each of the skeleton forms:

S20. $p\&\bar{q}$ **$p \vee q$

S21. $p\&q$ **$p \supset q$

S22. $\bar{p}\&\bar{q}$ **$p \supset q$

Both the logical and the "opposite direction" responses of the untrained Ss seemed to be inflated for these questions (43.8% and 35.5% respectively) at the expense of the '0' responses (20.6%, N=504). A high proportion of "opposite direction" responses were recorded to S20 and S22 (44.9% and 38.7%), possibly due to bidirectional negative implication from the presence of "not q" in the premises. For the trained Ss, 48.2% of the responses were logical, 33.0% were in the opposite direction and 18.7% were '0' (N=112).

The following four skeleton questions required deductions which were not information reducing (the logical response to each is an affirmation of the skeleton probe):

S23. $\sim(pq)$ **$p \supset \bar{q}$

S24. $p \vee q$ **$\bar{p} \supset q$

S25. p q **$\bar{p}$ q

S26. (pq) **$\bar{p}$ $\bar{q}$

68.2% of the responses of the untrained Ss were logical, 22.1% were in the opposite direction and 9.7% were '0' (N=719). For the trained Ss, these figures were 73.2%, 11.7% and 15.0% (N=153). The performance is clearly better on these questions than on the information reducing ones.

In order to determine whether information reducing deduction are counterintuitive or rather merely psychologically invalid, skeleton questions 13 through 16 were included. These questions required Ss to deduce "p or q" from "p" as a substep in a larger, information generating deduction. An example of a P-question generated from skeleton 13 was the following:

> 21. If either Cambodia or Laos can remain neutral for another six months, President Nixon can end the war and win the peace.
> Laos can remain neutral for another six months.
> **President Nixon can end the war and win the peace.

85.2% of the responses of the untrained Ss were logical, indicating that they had made the information reducing substep, 6.9% were in the opposite direction and 7.9% were '0' (N=826). For the trained Ss, these figures were 92.5%, 1.1% and 6.5% (N=186). Clearly, the information reducing substep is psychologically valid.

Inconsistent and Valid Formulas. It is well known that the fact that any sentence follows from a contradiction is highly counterintuitive. Skeleton question 27 was designed to see whether Ss would deduce an arbitrary sentence from a contradiction. 93.4% of the untrained Ss responses (N=376) and 88% of the trained Ss responses (N=75) were '0' as expected. The deduction of an arbitrary sentence from a contradiction is as follows:

*(1)	$p\bar{p}$	
*(2)	$\sim(p\bar{p}) \supset q$	(1)
*(3)	$\sim(p\bar{p})$	Valid
*(4)	q	(2), (3)

This deduction might be counterintuitive for two reasons (a) the step from (1) to (2) is information reducing and (b) the valid formula (3) must be generated. That some Ss do not spontaneously generate valid formulas is evident in comparing the untrained Ss' responses to skeleton 11 with those to skeleton 12, Whereas only 31.8% were logical

S11. $p \supset q\bar{q}$ $**\bar{P}$

S12. $(p \supset q\bar{q}) \& \sim(q\bar{q})$ $**\bar{p}$

to skeleton 11 (9.0% "opposite", 59.2% '0', N=223), 73.0% were logical to skeleton 12 (10.6% "opposite", 16.4% '0', N=189).

Baseline Figures. Skeleton questions 36-47 were designed to give an indication of the type of responding which would be obtained when Ss were asked to make deductions which were assumed to be relatively easy and intuitive. Roughly 73% of the responses to these questions were the logical response; this proportion varied with question type with a range from 41.5% to 96.8%.

Conclusions. Truth functional and predicate paleologic were examined. The class distortion principle of Von Domarus was found to be extendable to n-place predicates. Truth functional paleologic seemed to consist of truth functional logic conjoined with an additional axiom. The resulting system is inconsistent internally, but in ways that generally do not arise when information-reducing deductions are avoided.

Appendix: Skeleton Questions and Responses to Them.

The following are the skeleton questions from which the test questions were generated. The responses to each question are given by a vector (a b c d) in which the a entry is the percent of responses which affirmed the skeleton probe (which may correspond to a 'T' or 'F' response, depending on the question involved), the b entry is the percent of responses which denied the probe, the c entry is the percent of '0' responses and the d entry is the total number of responses to the question. Where two groups of Ss differed substantially in response pattern from one another, they are listed separately with the abbreviated group name to the left of the vector (tr = trained, not labeled tr = untrained, M = male, F = female, L = logical, I = intuitive).

S1. $p \supset q$ $**q \supset p$

	a	b	c	d
	(39.8%	28.7%	31.4%	191)
ML	(21%	21%	58%	52)
FL	(38%	31%	31%	42)
MI	(46%	29%	25%	48)
FI	(55%	51%	10%	49)

S2. $(p \supset q) \& (q \supset r)$ $**r \supset p$

	a	b	c	d
	(64.5%	15.1%	20.4%	152)
ML	(42%	13%	45%	38)
FL	(8%	9%	9%	33)
MI	(55%	23%	23%	40)
FI	(81%	15%	5%	41)

S3. $(p \supset q) \& (\bar{q} \supset r)$ $**p \supset \bar{r}$

	a	b	c	d
	(44.6%	27.4%	27.8%	186)
ML	(27%	18%	55%	49)
FL	(40%	31%	29%	42)
MI	(51%	32%	17%	47)
FI	(60%	29%	10%	48)

S4. p q $**\bar{p}$ $\bar{q}$

	a	b	c	d
	(44.4%	22.8%	32.8%	189)
ML	(23%	8%	69%	)
FL	(45%	24%	31%	42)
MI	(46%	29%	25%	48)
FI	(63%	29%	8%	51)

S1 - S4

tr(31.0% 31.0% 38.0% 158)

S5. p q ** (p $\bar{q}$)

(85.8% 10.1% 4.0% 375)
tr(95% 4% 1% 84)

S6. (p q) **p$\bar{q}$

(59.2% 20.3% 20.5% 409)
L(60.3% 14.8% 24.8% 189)
I(58.2% 25.0% 16.8% 220)
tr(48% 8% 44% 85)

S7. (p q) **p

(9.9% 8.1% 82.1% 223)
L(5.7% 3.8% 90.5% 106)
I(13.7% 12.0% 74.4% 117)

S8. (p q) **$\bar{q}$

(22.2% 10.1% 67.7% 189)
L(16% 8% 76% 91)
I(28% 12% 60% 98)

S7&S8.

tr (11% 2% 87% 84)
trL(0% 0% 100% 54)
trI(30% 7% 63% 30)

S9. (p$\vee$ q)$\supset$ r **$\sim$(p$\supset \bar{r}$)
(88.2% 8.0% 3.7% 187)

S10. [(p$\vee$ q)$\supset$ r]&(r$\supset$ s) **$\sim$(p$\supset \bar{s}$)
78.9% 10.2% 10.9% 128)

S9&S10.

tr(94% 4.5% 1.5% 66)

S11. p$\supset$(q$\bar{q}$) **$\bar{p}$

(31.8% 9.0% 59.2% 223)
tr(56% 2% 42% 41)

S12. $(p \supset (q\bar{q}) \& \sim (q\bar{q})$ **$\bar{p}$

(73.0% 10.6% 16.4% 189)
tr(68% 20% 12% 41)

S13. $[(p \vee q) \supset r] \& p$ **r

(86.3% 3.7% 10.0% 190)

S14. $(p \vee q) \supset r$ **$p \supset r$

(91.0% 6.4% 2.7% 188)

S15. $[(p \vee q) \supset r] \& (r \supset s) \& p$ **s

(79.5% 8.5% 12.0% 258)

S16. $[(p \vee q) \supset r] \& (r \supset s)$ **$p \supset s$

(86.3% 8.4% 5.2% 190)

S13-S16.

tr(92.5% 11.0% 6.5% 186)

S17. p **$p \vee q$

(50.2% 36.7% 13.0% 185)
M(66% 20% 14% 95)
F(33% 54% 12% 90)

S18. p **$q \supset p$

(27.4% 12.1% 60.5% 248)

S19. p **$\bar{p} \supset q$

(6.4% 10.9% 82.7% 156)

S170S19.

tr(41.6% 7.2% 51.1% 137)

S20. $p\bar{q}$ **$p \vee q$

(46.2% 44.9% 8.9% 158)

S21. pq **$p \supset q$

(45.8% 21.9% 32.3% 155)

S22. $\bar{p}$ $\bar{q}$ **$p \supset q$

(40.3% 38.7% 20.9% 191)

S20-S22.

tr(48.2% 33.0% 18.7% 112)

S23. (pq) **$p \supset \bar{q}$

(72.4% 22.8% 4.8% 189)

S24. $p \vee q$ **$\bar{p} \supset q$

(76.4% 15.5% 8.0% 187)

S25. $p \supset q$ **$\bar{p} \vee q$

(55.3% 27.1% 17.6% 188)

S26. $\sim(pq)$ **$\bar{p} \vee \bar{q}$

(68.3% 23.2% 8.4% 155)

S23-S26.

tr(73.2% 15.0% 11.7% 153)

S27. $p\bar{p}$ **q

(3.5% 3.1% 93.4% 376)
tr(12% 0% 88% 75)

S28. F(x)&G(x)&F(y) **G(y)

(29.9% 7.2% 62.8% 234)
ML(12% 2% 86% 66)
FL(31% 5% 63% 55)
MI(38% 8% 54% 61)
FI(42% 15% 42% 52)
tr(23% 2% 75% 60)

S29. F(x)&G(x)&$\bar{F}$(y) **$\bar{G}$(y)

(22.4% 4% 73.6% 125)
ML(3% 0% 97% 32)
FL(3.5% 3.5% 93% 28)
MI(34.5% 3% 62.5% 32)
FI(44% 12% 44% 34)
tr(7% 0% 93% 28)

S30. F(x,y)&G(x,y)&F(a,b) **G(a,b)

(23.9%	4.0%	72.1%	201)
ML(6%	0%	94%	48)
FL(13%	3%	85%	39)
MI(29%	4%	67%	55)
FI(41%	8%	51%	59)
tr(22.5%	2.5%	75%	40)

S31. F(x,y)&G(x,y)&$\overline{F}$(a,b) **$\overline{G}$(a,b)

(24.1%	8.9%	67.1%	158)
ML(2.5%	2.5%	95%	39)
FL(20%	74%	6%	35)
MI(37%	7%	56%	41)
FI(36%	45%	18%	44)
tr(21%	9%	70%	33)

S32. F(x,y)&F(a,b)&G(x) G(a)

(21.2%	6.5%	72.3%	184)
ML(5%	5%	90%	43)
FL(9.5%	9.5%	81%	42)
MI(31%	2%	67%	48)
FI(35%	10%	55%	51)
tr(14%	0%	86%	42)

S33. F(x,y)&$\overline{F}$(a,b)&G(x) $\overline{G}$(a)

(14.4%	7.7%	80.9%	209)
ML(2%	0%	98%	57)
FL(11%	4%	85%	47)
MI(29%	5%	65%	55)
FI(16%	10%	74%	50)
tr(20%	0%	80%	51)

S34. F(x)&F(a)&G(x,y) G(a,b)

(11.5%	8.3%	81.2%	252)
ML(6%	2%	92%	62)
FL(8%	6%	86%	64)
MI(14%	11%	76%	66)
FI(18%	15%	67%	60)
tr(13.5%	2%	84.5%	52)

S35. $F(x)\&\overline{F}(a)\&G(x,y)\ \ \overline{G}(a,b)$

(4.8%	2.4%	92.9%	126)
ML(0%	3%	97%	32)
FL(0%	0%	100%	28)
MI(6%	0%	94%	32)
FI(16%	6%	88%	32)
tr(7%	0%	93%	28)

S36. $p \supset (qr)$ $**\overline{q} \supset \overline{p}$

(72.7% 11.5% 15.8% 139)

S37. $p \vee q$ $**(pq) \vee (p\overline{q}) \vee (\overline{p}q)$

(64.9% 32.5% 2.6% 191)

S38. $(p \supset q)\&(q \supset r)$ $**p \supset r$

(91.0% 2.6% 6.4% 156)

S39. $(p \supset q)\&p$ $**q$

(96.8% 3.2% 0% 188)

S40. $p \supset q$ $**\overline{q} \supset \overline{p}$

(63.5% 17.9% 18.6% 156)

S41. $p \supset (s \vee r)$ $**s \supset (p \vee \overline{p})$

(50.0% 36.2% 13,8% 188)

S36-S41.

(72.9%	18.1%	9.0%	1018)
tr(7.18%	19.7%	8.4%	238)

S42. $F(a)\&G(b)\&G(c)$ $**F(c)$

(7.9% 6.9% 85.1% 202)

S43. $R(a)\&S(b)\&\overline{S}(c)$ $**\overline{R}(c)$

(11.1% 6.9% 82.0% 189)

S44. $p \supset q$ $**pq$

(70.5%	6.8%	22.6%	190)
tr(52%	7%	41%	42)

S45. $p \supset q$ **q

(12.7% 13.8% 73.5% 189)

S46. $[(pq) \supset (s \vee r)] \& (s \supset q)$ **$p \supset q$

(26.2% 32.3% 41.5% 195)

S47. $R(x,y) \& S(x,z) \& R(a,b)$ **$S(b,a)$

(11.2% 6.7% 82.0% 178)

S42&S43&S44&S45&S46&S47.

(13.9% 13.4% 72.7% 953)
tr(7.0% 7.4% 85.6% 215)

References

Abelson, R., & Rosenberg, M., Symbolic Psycho-logic: A Model of Attitudinal Cognition, Behavioral Science, 1958, 3, 1-13.

Amarel, S. On representations of problems of reasoning about actions. D. Michie (Ed.) Machine Intelligence 3. Edinburg University Press, 1968.

Arieti, S., Interpretation of Schizophrenia, New York: Bruner, 1955.

Bower, G. H. Organizational factors in memory. Cognitive Psychology, 1970, 1, 18-46.

Chapman, L., & Chapman, J., Atmosphere Effect reexamined, Journal of Experimental Psychology, 1959, 58, 220-226.

Collins, A. M. & Quillian, M. R. How to make a language user. In E. Tulving and W. Donaldson (Eds.) Organization and memory. New York: Academic Press, in press.

Davis, J. & Restle, F. The analysis of problems and prediction of group problem solving. Journal of Abnormal and Social Psychology, 1963, 66(2), 103-116.

Davis, J. The solution of simple and compound word problems. Behavioral Science, 9(4), 1964.

Dawes, R., Illogical Reasoning and Related Thinking Disorders, unpublished doctoral dissertation, University of Michigan, 1962.

Duncker, K. On problem solving. Psychological Monographs, 1945, 58, (Whole No. 270), 5.

Ellis, W. Memory of physically identical elements in human maze learning: A transfer problem. Psychological Bulletin, 1939, 36, 545-546.

Gagne, R. M. Human problem solving: internal and external events. In B. Kleinmuntz (Ed.), Problem solving: research, method, and theory. New York: Wiley, 1966. Pp. 127-148.

Gottesman, L. & Chapman, L., Syllogistic Reasoning Errors in schizophrenia, Journal of Consulting Psychology, 1960, 24, 250-255.

Haygood, R., & Bourne, L., Attribute and rule learning aspects of cognitive behavior, Psychological Review, 1965, 72, 175-195.

Janis, I. & Frick, F., The Relationship Between Attitudes toward Conclusions and Errors in Judging the Logical Validity of Syllogisms, Journal of Experimental Psychology, 1943, 33, 73-77.

Katona, G. Organizing and memorizing: Studies in the psychology of learning and teaching. New York: Columbia University Press, 1940.

Lucas, J. R., Minds, machines and Gödel. In A. R. Anderson (Ed.), Minds and Machines, Englewood Cliffs, N.J.: Prentics Hall, 1964, 43-59.

Meyer, D. E. On the representation and retrieval of stored semantic information. Cognitive Psychology, 1970, 1, 242-299.

Paige, J., & Simon, H. Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), Problem solving: research, method, and theory, New York: Wiley, 1966. Pp. 51-118.

Restle, F., & Davis, J. Success and speed of problem solving by individuals and groups. Psychological Review, 1962, 69(6), 520-536.

Shepard, R. N. & Chipman, S. Second-order isomorphism of internal representations: shapes of states. Cognitive Psychology, 1970, 1, 1-17.

Thomas, J. T., Jr. An analysis of behavior in the hobbits-orcs problem. The University of Michigan: Human Performance Center Technical Report, No. 31. August, 1971.

Tulving, E. Subjective organization and effects of repetition in multi-trial free recall learning. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 193-197.

Von Domarus, E., The Specific Laws of Logic in Schizophrenia. In J. S. Kasanin, (Ed.), Language and Thought in Schizophrenia, Los Angeles: University of California Press, 1944, 104-114.

Wason, P., Reasoning about a rule, Quarterly Journal of Experimental Psychology, 1968, 20, 273-281.

Wertheimer, M. Productive thinking, New York: Harper and Row, 1959.

Wilks, S. S., Mathematical Statistics, New York: Wiley & Sons, 1962.

Woodworth, R. & Sells, S., An Amtmosphere Effect in Formal Syllogistic Reasoning, Journal of Experimental Psychology, 1935, 18, 451-460.